Foundations and Trends® in Information Retrieval

Fairness in Search Systems

Suggested Citation: Yi Fang, Ashudeep Singh and Zhiqiang Tao (2024), "Fairness in Search Systems", Foundations and Trends® in Information Retrieval: Vol. 18, No. 3, pp 262–416. DOI: 10.1561/1500000101.

Yi Fang
Santa Clara University
yfang@scu.edu

Ashudeep Singh
Microsoft
ashudeep.singh@microsoft.com

Zhiqiang Tao
Rochester Institute of Technology
zhiqiang.tao@rit.edu

now
the essence of knowledge
Boston — Delft

Contents

Fairness in Search Systems

Yi Fang[1], Ashudeep Singh[2] and Zhiqiang Tao[3]

[1]*Department of Computer Science and Engineering, Santa Clara University, USA; yfang@scu.edu*
[2]*Microsoft, USA; ashudeep.singh@microsoft.com*
[3]*School of Information, Rochester Institute of Technology, USA; zhiqiang.tao@rit.edu*

ABSTRACT

Search engines play a crucial role in organizing and delivering information to billions of users worldwide. However, these systems often reflect and amplify existing societal biases and stereotypes through their search results and rankings. This concern has prompted researchers to investigate methods for measuring and reducing algorithmic bias, with the goal of developing more equitable search systems. This monograph presents a comprehensive taxonomy of fairness in search systems and surveys the current research landscape. We systematically examine how bias manifests across key search components, including query interpretation and processing, document representation and indexing, result ranking algorithms, and system evaluation metrics. By critically analyzing the existing literature, we identify persistent challenges and promising research directions in the pursuit of fairer search systems. Our aim is to provide a foundation for future work in this rapidly evolving field while highlighting opportunities to create more inclusive and equitable information retrieval technologies.

Yi Fang, Ashudeep Singh and Zhiqiang Tao (2024), "Fairness in Search Systems", Foundations and Trends® in Information Retrieval: Vol. 18, No. 3, pp 262–416. DOI: 10.1561/1500000101.

1

Introduction

> Equals should be treated equally and unequals unequally.
>
> *Aristotle, 384–322 BC*

Search systems are ubiquitous across a wide array of platforms, from online information sources such as web search engines, e-commerce sites, and social media to sociotechnical systems encompassing admissions, housing, and employment platforms. They significantly influence the flow of information and transactions, dictating the content that gets consumed, the products purchased, employment decisions, and admissions processes. The impact of these systems extends to both sides of the spectrum: they serve not only consumers, such as web users, employers, purchasers, and admissions officials, who rely on them to make informed choices but also providers, such as content creators, sellers, job applicants, and media organizations, whose visibility and success are directly affected by how they are ranked and presented within these systems. This dual influence underscores the substantial role that search systems play in access to information, shaping economic opportunities,

and social mobility. In recent years, there has been a growing focus within the Information Retrieval (IR) community on the *fairness* of search systems. This concern centers around whether the resources and benefits provided by these systems are equitably distributed among the various individuals or entities they impact. There is also a scrutiny of whether these systems perpetuate or introduce harms, especially those that are distributed in ways that are considered unfair or unjust.

Reflecting on the evolutionary trajectory of retrieval models over the past few decades reveals a significant shift towards data and machine learning driven methodologies. Initially, IR systems relied primarily on ranking algorithms that utilized various heuristics, such as TF-IDF weighting, to determine the relevance between a query and a document. The idea of aggregating multiple signals into the ranking process without resorting to heuristic methods led to the learning-to-rank techniques in the 2000s (Liu *et al.*, 2009), which involved defining hand-crafted features that capture different notions of what constitutes a relevant match, with machine learning models then tasked with learning the optimal combination of these features from training data. Recent neural IR models further eliminated the need for manual feature design (Mitra and Craswell, 2017). The rise of large language models (LLMs) is expected to dramatically transform the field of IR through their remarkable capabilities in language understanding, generation, generalization, and reasoning (Zhu *et al.*, 2023). These models bring a new level of sophistication to responding to complex queries. With the evolution of search engines into predominantly data-driven AI systems, they are increasingly susceptible to data and algorithmic biases. These biases can significantly impact the fairness of search results, potentially disadvantaging certain groups of consumers or providers, or reinforcing stereotypes.

In this monograph, we provide an introduction to fairness in search systems, with the aim of offering a starting point for understanding the problem space, reviewing the body of existing research, and laying the groundwork for further exploration and study in this critical area. Our focus is primarily on the fairness of a search system in delivering results that meet a user's information needs as encoded in their queries. We address fairness-related biases and harms, rather than the wider

spectrum of issues that search systems might encounter, such as the propagation of misinformation.

1.1 History of Fairness in Search

The history of fairness research in search has evolved over several decades, reflecting a growing understanding of how these factors impact the user experience and the ethical implications of IR systems.

In the early years of IR, dating back to the 1960s and 1970s, the primary goal was to provide users with a list of documents that contained the queried keywords. Early IR systems did not incorporate sophisticated algorithms for ranking these documents, and as a result, search results often lacked the depth and relevance of modern search engines. However, interestingly, unfair rankings were discussed by Cooper and Robertson in the probability ranking principle work (Robertson, 1977), even though they did not use the term "fairness" as such (Hiemstra, 2023). It was revealed that unfair rankings may arise from blindly applying the principle without checking whether its preconditions are met.

The 1990s saw a significant expansion in search with the advent of the Internet. The focus started shifting towards improving search algorithms for better relevance and precision. Google's PageRank algorithm (Page *et al.*, 1998) revolutionized search technology, which considered not only keywords but also the quality and relevance of web pages. As the commercial interests grew, search advertising became prominent. Advertisers could pay to have their content displayed when specific keywords were searched. This practice had the potential to introduce bias in search results, as the presence and ranking of content became influenced by commercial interests rather than purely by relevance and quality.

During this era, the aspects of diversity and novelty in search results began to gain attention, particularly in the context of providing a broad range of search results to users (Clarke *et al.*, 2008). As search engines became integral to daily life, concerns regarding bias in search results also began to surface. Algorithmic bias became a topic of discussion, especially as it related to the ranking of websites. Critics have argued that search engine algorithms sometimes favor authoritative sources

while marginalizing smaller or less mainstream voices in search results, in effect leading to concerns about information monopolies (Segev, 2010).

Discussions about net neutrality in the late 2000s and early 2010s also brought search engine neutrality into the spotlight, as part of the broader debate about equal access to online information (Crane, 2011). Search engine neutrality refers to the idea that search engines should have no inherent biases in their algorithms and should treat all web pages and content sources equally without favoritism. The central question was whether search engines should serve as neutral platforms that provided unfiltered and uncurated search results. The discussions about neutrality raised complex questions about the role of search engines as information gatekeepers and the potential consequences of curating content. Search engine providers faced increased scrutiny from regulatory bodies. They were challenged on practices such as favoring their own services in search results, penalizing competitor websites, and lack of transparency in their ranking algorithms. Legal battles and antitrust investigations became more common, as seen in the European Commission Guidelines on Ranking Transparency (Commission, 2020), as governments sought to ensure that search engines operated fairly and did not abuse their market dominance.

In the realm of IR research, numerous early studies have shed light on various forms of unfairness in search results. These encompass a range of biases, including racial, gender, and political viewpoint biases, which have raised concerns about the perpetuation of stereotypes through biased search outcomes. This area of inquiry is part of the broader research landscape focusing on fairness in sociotechnical and AI systems (Mitchell *et al.*, 2021), yet IR systems present their unique challenges and opportunities (Ekstrand *et al.*, 2022). Early work (Friedman and Nissenbaum, 1996; Introna and Nissenbaum, 2000) recognized the inherent capacity of search engines to incorporate social, political, and moral values into their ranking algorithms. To quantify the impact of such embedded values, Mowshowitz and Kawaguchi (2002) proposed a metric for measuring a search engine's deviation from an ideal exposure of content. Beyond the study of bias in algorithmic ranking, Vaughan and Thelwall (2004) and Vaughan and Zhang (2007) discovered that biases can arise from skewed crawling and indexing processes. Furthermore,

the concept of document retrievability (Azzopardi and Vinay, 2008) investigated the distribution skew in document retrievability across various retrieval systems, contributing valuable insights into the mechanics of search engine fairness.

In the 2020s, calls for ensuring fairness in search engine algorithms have intensified. Many raised concerns about the biases of AI and machine learning algorithms used in search engines (Baeza-Yates, 2018; Gao and Shah, 2020). The need to make these algorithms more equitable gained prominence. Ethical considerations became essential to the development and deployment of search engine algorithms. The relationship between the relevance of search algorithm results (and consequently, the revenue of the search engine) and the fairness of those results is not inherently contradictory. It has been shown that there are instances where enhancing the quality of the results, quantified by metrics such as Reciprocal Rank (RR), Average Precision (AP), or Normalized Cumulative Discounted Gain (nDCG), can also simultaneously improve the fairness of the outcomes (Hiemstra, 2023).

Fairness in search engines remains a dynamic and evolving field. In recent years, there has been a generally increasing number of publications on fair search as shown in Figure 1.1. The scope of this survey covers more than 400 papers including the representative papers about fairness studies in AI and the papers about fairness in search published in the top IR related conferences and journals such as SIGIR, CIKM, WSDM, WWW, KDD, ICTIR, ECIR, RecSys, FAccT, FnTIR, TOIS, ACL, EMNLP, NAACL, AAAI, IJCAI, NeurIPS, ICML, as well as some of the outstanding arXiv papers.

1.2 Fairness, Bias, and Diversity

While *fairness*, *bias*, and *diversity* are frequently discussed as interrelated concepts in the research community, their relationships remain complex and often misunderstood. According to the Cambridge Dictionary,[1] *bias* represents a disproportionate inclination for or against certain ideas or things, whereas *fairness* describes the equitable and reasonable

[1] https://dictionary.cambridge.org

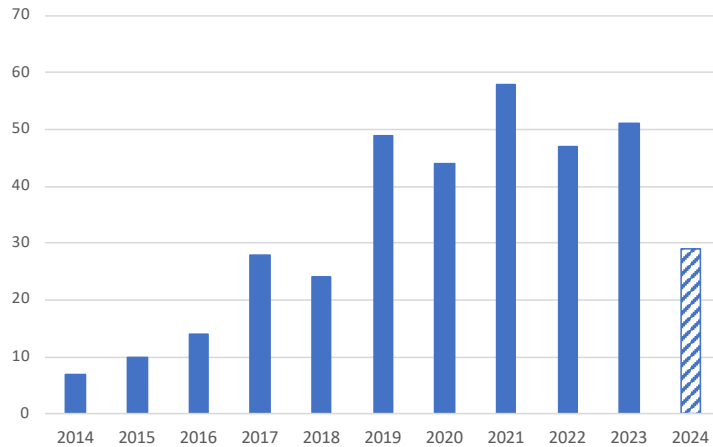

Figure 1.1: Publication trends in fairness in search (2014-2024). The data for 2024 is shaded to indicate that it represents an incomplete year at the time of this analysis.

treatment of individuals. This distinction is important: *bias* describes an observable characteristic of a system without making value judgments, while *fairness* addresses the ethical implications and societal impacts of system behavior (Ekstrand *et al.*, 2022).

Generally, different types of biases are key contributors to unfair outcomes in AI systems. The linkage between specific biases and resultant unfairness can be intricate (Li *et al.*, 2023). For instance, unfairness related to race and ethnicity might stem from biases in training data, model design, optimization algorithms, or evaluation benchmarks. Furthermore, a single type of bias, such as that in training data, can lead to various forms of unfairness such as individual and group unfairness.

On the other hand, the presence of bias does not inevitably lead to unfairness. For example, when a user searches for restaurants, a search engine shows results biased towards local establishments. This localization bias is based on the user's geographic location, which aligns with the user's likely intent. Beyond data and algorithmic biases, other factors can contribute to unfairness. It has been shown that certain fairness requirements are inherently conflicting, suggesting that upholding one type of fairness could inadvertently violate another (Kleinberg *et al.*, 2016).

Recent research in search systems has delved into various biases and debiasing methods (Zehlike *et al.*, 2022; Ekstrand *et al.*, 2022), but a clear distinction between research on bias and that on unfairness often remains elusive. Primarily, debiasing research tends to concentrate on enhancing retrieval performance, rather than explicitly promoting fairness. They usually conduct experiments based on improvements in relevance of results alone, using these gains to demonstrate the effectiveness of debiasing. In contrast, studies on fairness typically offer clear definitions and quantitative metrics for evaluating model unfairness, such as using performance disparities across groups to assess group-level unfairness. Fairness-focused research often assesses methods against both fairness metrics and traditional retrieval metrics.

While biases are recognized as key contributors to unfairness and debiasing methods can potentially improve fairness, many fairness studies do not rely on debiasing but instead directly incorporate fairness requirements into model design. This approach, like imposing fairness regularization during optimization, can sometimes compromise model accuracy. Hence, there is a discernible research gap between debiasing and fairness, despite their theoretical and practical interconnections (Li *et al.*, 2023). A more nuanced understanding of the relationship between bias, unfairness, and the interplay of debiasing and fairness enhancement methods could lead to more effective strategies that improve both fairness and accuracy in search systems.

Diversity in IR is about ensuring a wide range of information in search results. This means that the results should include a variety of sources, viewpoints, or content types, rather than being dominated by a few sources or perspectives. In many cases, efforts to improve fairness in IR systems also enhance diversity. For example, algorithms designed to reduce bias in search results often lead to a more diverse set of search results. On the other hand, there can be tensions between these two goals. For example, maximizing diversity in search results might sometimes lead to less fair outcomes for certain groups, or vice versa. In the literature, the notion of coverage-based diversity (Drosou *et al.*, 2017) is most closely related to fairness, which requires that members of multiple, possibly overlapping, groups, be sufficiently well-represented among the top-k, treated either as a set or as a ranked list. Both fairness

and diversity should consider the user perspective. An IR system might be fair and diverse from a content perspective but still fail to meet the diverse needs and fairness expectations of different user groups.

Fairness is frequently encapsulated within the broader framework of FACTS-IR that stands for Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval that also contains the other pivotal aspects of responsible IR. The report from the FACTS-IR Workshop (Olteanu *et al.*, 2021) delves into the interplay and significance of these concepts. In this survey, our primary focus is on fairness, although we will also touch upon the other aspects, particularly in contexts where they intersect with or influence fairness.

1.3 Biases in Search

The search process can be conceptualized as a feedback loop encompassing various stages, such as query formulation and understanding, document representation, retrieval (or candidate generation), ranking, user feedback, and evaluation. At each of these stages, biases may arise, and the cyclic nature of the feedback loop has the potential to sustain or even intensify these biases. While this survey primarily focuses on fairness, it is important to recognize that various types of biases are significant contributors to unfair outcomes in search systems. A thorough understanding of how these biases interplay is essential for delivering fair and accurate information to users. In this section, we outline the architecture of a typical search engine, highlighting potential biases at each stage as depicted in Figure 1.2. While this list of biases is not exhaustive, it aims to provide an initial understanding of how biases can manifest throughout the search process. More detailed discussions on biases and unfairness, their implications, and mitigation strategies are provided in the subsequent sections.

Given data sources, **crawling and indexing** are the foundational processes in search engines that determine what content becomes searchable. Crawling is the first step where crawlers, also known as spiders, systematically browse the web to collect data from accessible web pages. Due to the extensive nature of the web, *crawling bias* may occur when these crawlers favor certain pages over others based on factors such as

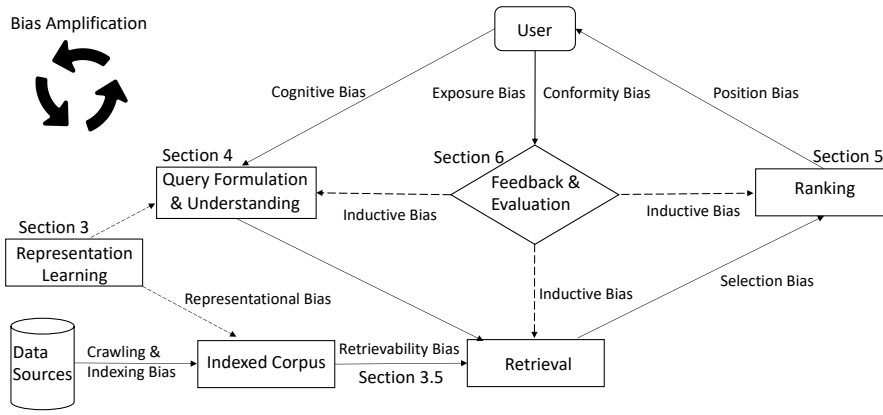

Figure 1.2: An overview of biases that can emerge at various stages in the life cycle of a search system. Section x in the figure refers to the specific section where the corresponding fairness issues are discussed.

page popularity or the quality and quantity of incoming links. This prioritization can result in the underrepresentation of less popular or newly established websites. Additionally, *indexing bias* can arise during the organization and storage of data, where a search engine might prioritize certain content, potentially distorting representation based on aspects like language, popularity, or perceived relevance. This can disproportionately represent cultural and linguistic content. Moreover, technical constraints and operational guidelines, such as the use of robots.txt files to guide crawler activities, can inadvertently introduce biases.

Query formulation and understanding begin with the user entering a query into the search engine. It involves a multi-faceted analysis of user queries to interpret their intent, context, and meaning. A *cognitive bias* is a systematic pattern of deviations in thinking which may lead to errors in judgments and decision-making (Azzopardi, 2021). Such biases may significantly influence how users formulate their queries. For instance, *confirmation bias* stems from people's tendency to prefer confirmatory information, where they discount information that does not conform to their existing beliefs. When querying, this may manifest as people employing positive test strategies where they try to find information that supports their hypotheses.

Representation learning involves transforming documents or queries into a format that can be efficiently processed by a search system. During this stage, each document/query is analyzed and converted into a structured form, often as a vector of features, which is then indexed and stored in the search system's database. This process also involves pre-processing steps such as tokenization, removal of stop words, and stemming or lemmatization. The goal is to distill the essence of each document into a representation that captures its main themes and content in a way that can be readily compared with user queries, facilitating effective and efficient retrieval in response to search requests.

Representational bias may emerge in representation learning. This bias can stem from a variety of factors related to the content, sources, and historical context of the documents. It manifests as skewed or unbalanced perspectives, representations, or information within the corpus itself, which can lead to a misrepresentation of certain demographics, viewpoints, or subject areas, affecting the fairness and accuracy of the search process. *Representational bias* is not introduced by the retrieval algorithms but rather originates from the intrinsic characteristics of the corpus. Bias inherent in training corpora can not only persist but also amplify (Papakyriakopoulos *et al.*, 2020; Wang *et al.*, 2024c) in learned latent representations through deep neural networks, such as pre-trained word embeddings (Brunet *et al.*, 2019), BERT (Kurita *et al.*, 2019), and more recently in LLMs (Gallegos *et al.*, 2023).

Retrieval is a process that retrieves all the candidates that match the user query from the index. In general, the retrieval system has to be fast and lightweight, as it considers the contents of the entire index. *Retrievability bias* measures how easily a document can be retrieved and exposed to the later ranking stage. A system with pronounced *retrievability bias* disproportionately favors certain documents over others (Azzopardi and Vinay, 2008), potentially resulting in unfair outcomes in the search results (Otterbacher *et al.*, 2017). *Popularity bias* can also be manifested in retrieval, which is the tendency to retrieve popular items more frequently than their intrinsic popularity justifies. This bias stems from several contributing factors. The sheer volume and visibility of content from popular sources can overshadow less popular but relevant content in the retrieval process. Many search engines use

link analysis algorithms such as PageRank to infer its importance or relevance. Popular pages with many inbound links are more likely to be retrieved due to their perceived authority. Some retrieval algorithms may use historical user interaction data, like click-through rates as indicators of relevance. Popular items that have been clicked on or interacted with more frequently are likely to be considered more relevant, thus being retrieved more often.

Ranking involves reordering the top results obtained from the retrieval process. This can be based on chronological order, relevance criteria, or a combination of both. Learning-to-rank techniques are often employed at this stage to enhance the relevance of the results (Liu *et al.*, 2009). Beyond the *popularity bias* noted in the retrieval stage, the ranking stage is also subject to biases introduced during retrieval. Specifically, *selection bias* occurs when the initial set of documents retrieved dictates the subsequent ranking order (Wang *et al.*, 2023c). If this initial retrieval is biased or narrow in scope, the range of documents available for ranking becomes limited. As a result, the ranking stage is constrained to working with this pre-selected set, potentially overlooking more relevant or diverse documents that were not initially retrieved.

When ranked results are presented to the users, *position bias* occurs when users engage more frequently with items at the top of a ranked list, often irrespective of the actual relevance of these results. Eye-tracking studies have shown that users typically focus on the initial items and are less likely to consider those positioned lower (Joachims *et al.*, 2007b). Other research indicates that users often place undue trust in the top-ranked results and may not evaluate subsequent items as thoroughly, leading to a lack of holistic assessment of all available results (O'Brien and Keane, 2006).

User feedback on ranked search results can be categorized into two types: explicit and implicit. Explicit feedback is provided directly by users in a clear and intentional manner such as ratings and surveys. It represents a deliberate effort to convey relevance satisfaction with the search results. Explicit feedback can also be done by third-party human annotators by providing relevance judgment on query-document pairs. Implicit feedback is gathered from user behavior and interactions that are not directly intended as feedback but can be interpreted

as such. It is unobtrusively collected as users go about their normal activities. Examples include click-through rate (CTR), dwell time, scroll depth, mouse movements, query reformulations, bounce rate, and so on. **Evaluation** is required to continuously monitor the performance of a search engine, as well as for measuring the effect of new changes that are introduced to any of its components. Evaluation can be done either manually, using explicit feedback, or automatically by tracking the implicit feedback such as clicks and session metrics.

Conformity bias can skew user explicit feedback, as individuals often align their behaviors with group norms, sometimes overriding their personal judgment (Azzopardi, 2021). This can lead to feedback that does not accurately represent their true opinions. Similarly, *confirmation bias* occurs when users selectively favor or emphasize search results that align with their pre-existing beliefs. This bias can result in feedback that reflects personal preferences or beliefs rather than an impartial assessment of the search results' quality.

Unlike explicit feedback, implicit feedback only offers a limited indication of user preference, as it lacks accurate information on what users like or dislike. *Exposure bias* is a significant issue in this context, arising from the fact that users only interact with a subset of documents. Consequently, not all unobserved interactions imply a negative preference. This ambiguity stems from two potential reasons for an unobserved interaction: either the document was not relevant to the user, or the user was simply unaware of it. This makes it challenging to accurately differentiate between genuinely negative interactions where the user is exposed to but not interested in a document and potentially positive ones where the user is not exposed to the document. As a result, this inability to distinguish between different types of unobserved interactions can lead to substantial biases in the learning process (Chen *et al.*, 2023b).

User feedback and evaluations are pivotal to update the parameters of machine learning models in various components, including query understanding, retrieval, and ranking, thus creating a feedback loop. To enhance specific desirable properties, *inductive biases* can be intentionally incorporated into the model design. *Inductive biases* are the underlying assumptions that a model uses to better learn the target

function and generalize beyond the training data. These biases are often not harmful but essential, as the core of machine learning is the ability to extrapolate predictions to new, unseen examples. Without making certain assumptions about the data or model, generalization is impossible, as the output for unseen examples could vary widely. The development of an effective search system requires the incorporation of specific assumptions about the nature of the target function to guide the learning process. Moreover, some unfairness mitigation strategies, such as the in-processing methods discussed in Section 5, leverage inductive bias to correct for certain biases.

As shown in Figure 1.2, the search process forms a feedback loop and biases emerge in different stages of the loop. These biases could be further amplified over time along the loop. Take *popularity bias* or *position bias* as an example. Initially, certain documents may be ranked higher due to their popularity or early user engagement. These documents then garner additional feedback, which influences future rankings, potentially fostering a rich-get-richer dynamic (Joachims *et al.*, 2017c). This phenomenon raises important fairness questions regarding how exposure should be distributed, ideally based on the merit of the documents or items, rather than their initial popularity or position (Biega *et al.*, 2018; Singh and Joachims, 2018). For instance, in a job applicant ranking system, such dynamics could exacerbate existing unfairness, such as gender disparities. Similarly, in an online marketplace, this bias could favor certain sellers (or groups), leading to monopolistic tendencies and potentially driving other sellers out of the market (Morik *et al.*, 2020). Both scenarios highlight the important need to address the biases and feedback loop to prevent the reinforcement of existing disparities in search systems.

1.4 Comparisons with Related Surveys

In recent years, a number of surveys discussing fairness and bias in general machine learning have been published (Caton and Haas, 2020; Castelnovo *et al.*, 2022). They usually focus on the fairness works in classification tasks. A few surveys provide an overview of fairness in recommendation tasks (Wang *et al.*, 2023b). Recommendation algorithms

can usually be considered as a type of ranking algorithm, but they often represent different characteristics. Pitoura *et al.* (2021) addresses fairness in both ranking and recommendation, and Ekstrand *et al.* (2022) discusses fairness in information access systems such as information retrieval and recommendation. Chen *et al.* (2023b) provides a survey on bias and debias in recommender systems, which covers a part of the content about fairness in recommendation. Similarly, Li *et al.* (2023) offers a systematic survey of existing works on fairness in recommendation by focusing on the foundations for fairness in recommendation literature. Recently, Dai *et al.* (2024a) presents a survey on bias and unfairness in IR systems that incorporate large language models predominantly references studies from the recommendation systems domain. While covering a brief introduction about fairness in classification and ranking, our survey pays specific attention to organizing the concept of fairness in search through a comprehensive taxonomy of fairness notions proposed in search problems, the task-specific techniques for promoting ranking fairness, as well as the datasets specially suitable for fairness research in search.

Three surveys were focused on fairness in ranking and retrieval systems (Ekstrand *et al.*, 2022; Zehlike *et al.*, 2022; Patro *et al.*, 2022). One recent survey performed a systematic literature review of the field of fairness, accountability, transparency, and ethics in information retrieval (Bernard and Balog, 2023). Our survey distinguishes itself from existing literature by offering several key advantages: 1) it provides a holistic review of unfairness across the entire life cycle of a search process, in contrast to previous surveys that primarily concentrate on fairness in ranking; 2) it introduces a thorough taxonomy of fairness in search and retrieval, aiding readers in comprehending various fairness considerations within search systems and facilitating an organized framework for navigating the literature in this domain; and 3) it is designed to be accessible, enabling newcomers to the field to develop a systematic understanding of the subject.

It is also worth noting that there have been several tutorials and workshops related to investigating biases and fairness issues in IR including the following: Addressing Bias and Fairness in Search Systems at SIGIR 2021 (Gao and Shah, 2021), Fairness of Machine Learning in

Recommender Systems at CIKM 2021 (Li *et al.*, 2021b), Fair Graph Mining at CIKM 2021 (Kang and Tong, 2021), Gender Fairness in Information Retrieval Systems at SIGIR 2022 (Bigdeli *et al.*, 2022), Fairness of Machine Learning in Search Engines at CIKM 2022 (Fang *et al.*, 2022), Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era at KDD 2024 (Dai *et al.*, 2024a) and WSDM 2025, and the workshop series on Algorithmic Bias in Search and Recommendation (BIAS) at ECIR 2020-2023 (Boratto *et al.*, 2023) and SIGIR 2024 (Bellogin *et al.*, 2024).

1.5 Intended Audience and Scope

This survey is beneficial for a wide array of individuals in the information retrieval field, including: 1) newcomers seeking a comprehensive guide to quickly delve into fairness issues in search systems; 2) those grappling with various sources of bias and requiring a systematic study to grasp the nuances of unfairness in search; 3) researchers aiming to stay up-to-date with cutting-edge techniques for mitigating unfairness in search; and 4) practitioners confronting unfairness challenges in the development of search systems and searching for effective solutions.

Primarily written for the IR community, this monograph also caters to diverse backgrounds such as machine learning, natural language processing, and AI ethics. It serves as an accessible entry point to the concept of fair search, enriched with numerous practical insights. We envision this resource as valuable for students, researchers, and software practitioners alike. Offering a holistic perspective and a thorough exploration of key ideas, it is essential for understanding and constructing modern search systems. These systems are crucial in enabling billions of users to access a wealth of global knowledge and services while ensuring fairness and equity in access.

1.6 Structure of the Survey

The monograph is structured as follows.

- Section 1 describes the architecture of a modern search system with important components and highlights various biases that

may arise in the search process. We also briefly review the history of fairness in search.

- Section 2 provides background information about the bias in algorithmic decision-making in general and in search in particular. We review the existing work on bias mitigation in machine learning and discuss the challenges in this space.
- Section 3 focuses on representation learning and content analysis, and on how to learn an unbiased data representation.
- Section 4 investigates fairness in query understanding, specifically in query formulation, query suggestion, and non-textual queries.
- Section 5 studies fair ranking and how to mitigate unfairness in rankings.
- Section 6 discusses bias in relevance judgment (both explicit and implicit) and how to learn and evaluate with biased feedback.
- Section 7 discusses emerging research directions, prompted by the rise of large language models (LLMs) and the growing imperative for responsible AI. This section also examines the open challenges that define this evolving landscape.

2

Background and Foundation

As machine learning finds widespread applications in the world around us, especially in areas such as healthcare, public policy, and law enforcement, there is significant interest in understanding the societal impact of these systems. Although it is a common belief that algorithmic decisions (e.g., those based on statistical modeling and machine learning) can counteract some existing biases and inconsistencies in human decision-making, data-driven decision-making also affords new mechanisms to introduce unintended bias (Barocas and Selbst, 2016). Some prominent studies have brought fairness issues into the limelight. For example, in 2016, ProPublica published a study on COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) that was designed to help human judges make bail decisions in the criminal justice system (Angwin and Larson, 2016a; Angwin and Larson, 2016b). They found that the false positive rate of the risk score model was significantly higher for black defendants. Later, in 2018, Buolamwini and Gebru (2018) showed that commercially available face recognition systems were highly inaccurate for faces with darker skin tones, and this was also representative of the diversity of the dataset, which consisted primarily of lighter skin tone faces. These studies highlight several complex ideas

of bias and fairness in data-driven decision-making systems, that we will try to describe in this section.

In this section, we will lay a foundation for the work done on the sources of unfairness in algorithmic decision-making systems (Section 2.1), define notions of fairness for algorithmic decision-making (Section 2.2), describe various ways in the literature to mitigate unfairness with respect to these notions of fairness (Section 2.3), and discuss the application of machine learning fairness to IR systems (Section 2.4). In the rest of the monograph, we will rely on this foundation to extend these notions of fairness to aspects of a search system.

2.1 Sources of Unfairness in Machine Learning Systems

Unfairness in machine learning models can arise at any stage of the model development process from the data pre-processing stage to evaluation and deployment. It can arise due to the biases already present in society, such as historical discrimination like redlining that has long-term effects on variables like wealth. Data collection processes can introduce unfairness through sampling biases, response biases, how variables are defined and measured, and what perspectives are captured in the data. During the learning process, machine learning models can also directly use sensitive attributes, or learn proxies for sensitive attributes, introducing unfairness. The objective functions and the models are also optimized for reflecting certain perspectives. Unfairness can arise in model evaluation if the same issues around input data biases also apply to evaluating that data. The choice of success metrics can also prioritize some stakeholders over others. Human response to model outputs is also a potential source of unfairness as humans may make inaccurate assumptions about the model and the world. Moreover, it is important to recognize that even if biases and unfairness are *removed* from any stage, they can re-emerge at any stage, and if not handled, they can also be propagated or reintroduced in the long term. Ideally, each source of bias requires different interventions and measurements.

2.1.1 Worldviews and Assumptions

Any attempt to design a fair decision-making mechanism has to make assumptions about how the data is observed. Friedler *et al.* (2021) proposed a framework to think about the underlying assumptions about fairness and the treatment of bias, as illustrated in Figure 2.1. Specifically, the authors discuss three key conceptual spaces: the construct space (CS) that represents the true, unobservable characteristics and qualifications of individuals, for example, an individual's innate intelligence or work ethic; the observation space (OS) that represents the quantified features we can actually measure about individuals, such as test scores or job performance ratings; and finally, the decision space (DS) that represents the decisions made about individuals, such as their rank in an ordering or rating. Although an OS serves as a proxy for the CS to map individuals to the DS, it may actually be biased relative to the CS, that is, it may not respect the ordering of individuals in the CS.

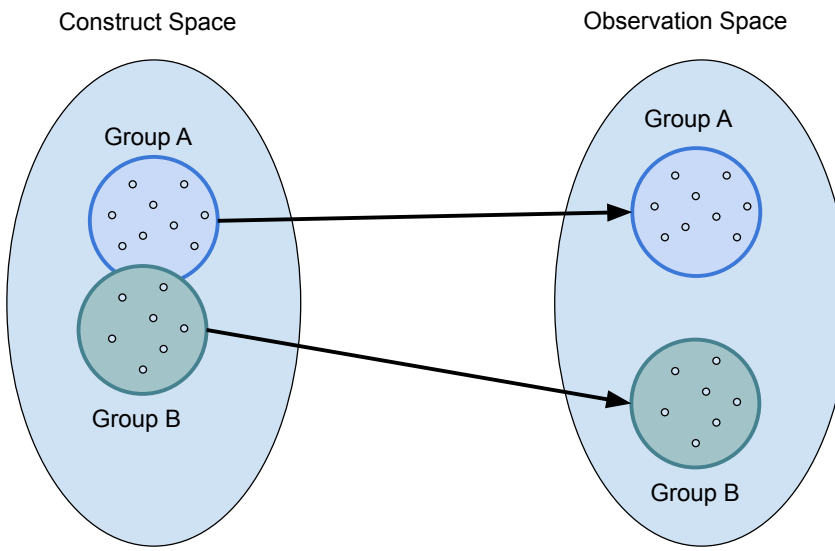

Figure 2.1: An illustration of the construct space and the observation space as described by Friedler *et al.* (2021). In each space, individuals in two different groups are represented. In this example, even if the two groups are closer in the construct space, they might appear farther in the observation space used to make decisions.

Given this framework of conceptual spaces for decision-making, two extreme worldviews can be defined with respect to algorithmic decision-making and, especially, machine learning-based decision-making systems. *What You See Is What You Get* worldview, often referred to as WYSIWYG, assumes that the OS (consisting of scores, qualifications, etc.) accurately reflects the true properties of individuals. Any differences seen between groups are taken at face value. Meanwhile, *We're All Equal* (WAE) worldview assumes that observed differences between groups are solely the result of biased processes or observations. It assumes that groups have equal distributions of qualifications and merit in the true underlying construct space. The choice of worldview and mitigation strategy depends on context and assumptions about the source of unfairness. However, making these assumptions explicit is crucial for selecting appropriate fairness evaluation methods and interventions. This work also argues that clarity on the normative underpinnings of fairness methods is currently lacking in much of the research, but is essential for understanding the effects of chosen fairness evaluations and mitigation strategies.

2.1.2 Harms: Distributional vs. Representational

Crawford (2017) described a framework of harms to describe the impact of unfair machine learning methods on individuals and groups, consisting of two categories: Distributional harms and representational harms. While distributional or allocative harms refer to harms caused by unfair distribution of resources or opportunities, representational harm refers to harms caused by biased or unfair representation and misrepresentation. For example, a group facing discrimination gets fewer opportunities in education, employment, loans, etc. would be considered distributional harm, a negative or stereotypical portrayal of a group in search results, media, or films that shape public perception would be considered representational harm. A study by Kay *et al.* (2015) detected the presence of gender bias in image search results for a variety of occupations, like doctor, nurse, teacher, etc. (e.g., Figure 2.2) and conducted a user study to emphasize how a biased information environment may affect users' perceptions and behaviors, by showing that such

Figure 2.2: Example of representational harm: An image search result page for the query "CEO" showing a disproportionate number of male CEOs.

biases indeed affect people's belief about various occupations. Distributional harm occurs at both an individual and group level, and can be measured by quantitative measures as well as remedied by focusing on the fair allocation of resources and opportunities. On the other hand, representational harm occurs only at a group level, is hard to evaluate quantitatively, and may be remedied by ensuring fair, accurate, and respectful portrayal which is often more challenging than remedying distributional harm.

2.2 Defining Fairness Notions

There are two paradigms for defining notions of fairness in machine learning. One considers fairness with respect to individuals (referred to as *Individual Fairness*), and the other considers fairness with respect to the groups that individuals belong to in terms of their sensitive attributes (referred to as *Group Fairness*).

2.2.1 Individual and Group Fairness

The individual fairness perspective states that two individuals similar with respect to a task should be classified similarly (Dwork *et al.*, 2012). Other examples include Counterfactual Fairness (Kusner *et al.*, 2017; Kilbertus *et al.*, 2017), and Treatment Equality.

Group fairness definitions can be divided into three major categories: Independence, Separation, and Sufficiency in terms of predictions $\hat{Y}$, true labels Y, and sensitive attributes A (Barocas *et al.*, 2019). While the notion of *Independence* between A and $\hat{Y}$ implies constraints such as De-

mographic Parity (Calders *et al.*, 2009), *Separation* implies conditional independence with respect to a sensitive attribute A, i.e., constraints such as Equalized Odds and Equal Opportunity (Hardt *et al.*, 2016). Meanwhile, the notion of *Sufficiency* implies that the predictor is well calibrated for all sensitive attributes (Pleiss *et al.*, 2017). In short, *Independence* ensures no correlation between sensitive attributes and predictions, *Separation* requires equal error rates across groups, and *Sufficiency* demands predictions be equally informative across groups for fair outcomes in machine learning.

Independence: Demographic Parity. Demographic parity is an independence-based notion of fairness that is satisfied if $\hat{Y}$ is independent of A, i.e., $\hat{Y} \perp A$. In other words, for two groups $A = a$ and $A = b$:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b).$$

Note that, in the definition above, the true labels Y are not involved in the definition of the fairness metrics which means that the constraint is not concerned with how accurate the model is, but only that the prediction rates be the same for two groups. This has certain shortcomings, for example, a perfect classifier ($\hat{Y} = Y$) does not always satisfy the constraint, for example, in the scenario when two groups have different rates of $Y = 1$ ground truth labels. In other words, if A and Y are correlated, a model that satisfies this constraint might need to significantly sacrifice accuracy. However, it must be noted that given a choice of labels Y, the prediction rates might be different for different groups because they are genuinely different (e.g., due to group preferences) or because the ground truth labels reflect certain historical biases. This also relates these fairness constraints to the discussion on Worldviews in Section 2.1. Several error-based notions of fairness alleviate such concerns, as we will discuss next.

Separation: Error-based Fairness definitions. A classification model's predictions can be summarized into a confusion matrix when the true labels are known (like in Table 2.1). Each cell of the confusion matrix leads to a metric that is often used to measure the types of errors that the model is making (like false positives) or getting right (like true

Table 2.1: A confusion matrix is used to define different metrics based on the types of errors made by a machine learning model. Equality among groups for each of these metrics defines error-based notions of fairness.

	Y=1	Y=0	$P(Y=1\|\hat{Y})$	$P(Y=0\|\hat{Y})$
$\hat{Y}=1$	True Positive	False Positive	$P(Y=1\|\hat{Y}=1)$ Positive Predictive value	$P(Y=0\|\hat{Y}=1)$ False discovery rate
$\hat{Y}=0$	False Negative	True Negative	$P(Y=1\|\hat{Y}=0)$ False omission rate	$P(Y=1\|\hat{Y}=0)$ Negative predictive value
$P(\hat{Y}=1\|Y)$	$P(\hat{Y}=1\|Y=1)$ True positive rate	$P(\hat{Y}=1\|Y=0)$ False positive rate		
$P(\hat{Y}=0\|Y)$	$P(\hat{Y}=0\|Y=1)$ False negative rate	$P(\hat{Y}=0\|Y=0)$ True negative rate		$P(Y=\hat{Y})$ Accuracy

positives). Equality among different groups on each of these metrics leads to a definition of fairness. For example, Hardt *et al.* (2016) define Equal Opportunity as the equality between true positive rates ($P(\hat{Y}=1|Y=1,A=a)=P(\hat{Y}=1|Y=1,A=b)$) which is also equivalent to $\hat{Y}\perp A|Y=1$. Alongside, if the equality between false positive rates (or true negative rates) is satisfied, the notion of fairness is called Equalized Odds (Hardt *et al.*, 2016) or Separation (Barocas *et al.*, 2019).

Sufficiency: Calibration-based Fairness definitions. However, since the predictions may not always be a binary variable, but a continuous value like a probability between 0 and 1, several calibration based notions of fairness can be defined. Such notions describe how likely an individual with a prediction of $\hat{Y}=p$ is to belong to the group $Y=1$ or $Y=0$. Gaps in how well a model is calibrated for different groups can be used to define fairness. In risk assessment tools, like COMPAS, one of the concerns raised is exactly how a *high* risk score for black defendants is not the same as a *high* score for other defendants.

2.2.2 Causal Definitions of Fairness

Several other techniques formulate the measurement and mitigation of disparities based on causal reasoning and intervention, rather than merely statistical relationships (Kilbertus *et al.*, 2017; Kusner *et al.*, 2017; Nabi and Shpitser, 2018). Using causal language (for example, structural models of the world) can make value judgments more explicit

and allow practitioners to designate different pathways through which correlations or dependencies between sensitive attributes and predictions/decisions may be acceptable or not. Overall, causal reasoning is a great way to design interventions that reduce disparities and improve overall outcomes.

2.3 Mitigating Unfairness in Machine Learning

Mitigation approaches to prevent unfairness in ML models can be generally categorized into three categories: Pre-processing, In-processing, and Post-processing (Caton and Haas, 2020). Each of these methods targets different stages of the ML pipeline, as illustrated in Figure 2.3.

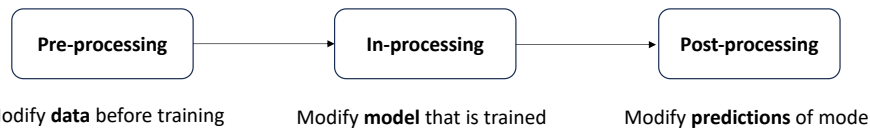

Figure 2.3: Mitigating unfairness at different stages of the machine learning pipeline.

Pre-processing methods stem from the understanding that biases in training and evaluation data are a primary source of unfairness in ML models. These biases can manifest in various forms, such as imbalanced data distributions, noise in input features, and inaccuracies in labels, particularly in relation to sensitive attributes. To mitigate this, the strategy involves pre-processing the training data to reduce these biases before the model training begins. The key idea is to train a model on a repaired dataset so that the model can be inherently fair. Pre-processing methods typically focus on modifying the sample distributions of protected variables or implementing specific data transformations with the objective of mitigating discriminatory biases from the training dataset.

Pre-processing methods consider fairness as the first concern in the ML pipeline. Their major advantage lies in their versatility: they are not tied to any specific modeling technique used later in the ML process. However, these methods can have unpredictable effects on the accuracy of the model and may not fully address unfairness in the test data.

Furthermore, there are scenarios where altering the training data is not permissible due to technical constraints or legal considerations, limiting the use of pre-processing methods in such cases.

In-processing methods are applied to the design and training of models, which can induce intrinsically fair models. The majority of the existing works of fairness in ML fall into this category. They often aim to balance the accuracy and fairness demands by modifying the learning process, e.g., incorporating fairness metrics into the objective function of the main learning task.

One of the main strengths of in-processing methods is their ability to offer more effective trade-offs between accuracy and fairness, as finding this balance is central to their design. A notable drawback is that such methods often lead to non-convex optimization problems, which do not guarantee optimal solutions. Furthermore, altering the learning process of a machine learning model may not always be feasible. Another complexity with in-processing methods is the ambiguity surrounding the underlying worldview of fairness-aware models, given their dual objective of balancing fairness with accuracy (Zehlike *et al.*, 2022).

Post-processing methods are implemented by applying transformations to a model's output in order to mitigate unfairness. Typically, this involves modifying the model's predictions to conform to specific fairness constraints, as seen in works like Hardt *et al.* (2016) and Kallus and Zhou (2019).

One of the main advantages of post-processing is its flexibility. It requires only access to the model's predictions and information about sensitive attributes, without the need to modify the underlying algorithms or machine learning models. This characteristic makes post-processing particularly suitable for scenarios where the ML pipeline operates as a black box and its internal workings are not fully accessible. Additionally, many post-processing methods can ensure a certain level of representation or visibility for protected groups. A general downside of post-processing is the implication that achieving fairness necessitates a compromise in accuracy. This is because it relies on adjusting the

outputs of an already trained model. Consequently, post-processing may inadvertently reinforce the notion that fairness and accuracy are mutually exclusive in some contexts.

2.3.1 Challenges and Limitations in Mitigation

There are fundamental trade-offs when it comes to satisfying multiple fairness criteria all at once, unless the classifier is perfect or the base rate of the true labels is the same (Kleinberg, 2018; Chouldechova, 2017). This means that the choice of what fairness criteria to satisfy must be chosen with careful consideration and with respect to the problem domain at hand.

2.4 Applying ML Fairness Definitions to IR

ML fairness notions, definitions, and mitigation methods discussed in this section do not directly translate to information retrieval systems such as search, primarily because search systems are often comprised of multiple stages (as discussed in Section 1.3 and Figure 1.2) as compared to prediction models that predict labels or scores for a given input. We will study these different stages from the lens of fairness in subsequent sections. Furthermore, the output of a search system is traditionally a ranked list of documents (websites, images, products, etc.) making the output space much richer than the predictions of classification models, simply because the rank of a document is not independent of the other documents in the candidate set. Based on this key difference, novel notions of fairness need to be developed for rankings which we will discuss in Section 5.

3

Representation Learning and Content Analysis

Representation learning (Bengio *et al.*, 2013) has been long studied and achieved great success in contemporary IR systems, especially since the rapid development of deep neural networks, covering a series of milestones, such as word/document embeddings (Le and Mikolov, 2014), sequence-to-sequence learning with RNNs (Sutskever *et al.*, 2014; Cho *et al.*, 2014), CNN-based methods (Kim, 2014; Severyn and Moschitti, 2015), and the recent transformer family (Vaswani *et al.*, 2017; Devlin *et al.*, 2019). In this section, we will first discuss the bias in different learned latent representations based on neural networks, particularly focusing more on recent document/query embedding methods and large language models. Then, we will elaborate on three mainstream representation learning methods (i.e., word embeddings, pre-trained language models, and multimodal embeddings) in the fair ranking context, each of which introduces preliminary knowledge, its connection and applications in ranking, and how they can help realize fairness. Finally, we discuss bias in the retrieval pipeline and contents.

3.1 Bias in Learned Latent Representations

This section investigates the bias rooted in learning representations with deep neural networks for the ranking problem, including bias in word embeddings (Papakyriakopoulos *et al.*, 2020; Brunet *et al.*, 2019), pre-trained language models (Wang *et al.*, 2022b; Rekabsaz and Schedl, 2020), and multimodal/crossmodal embeddings (Jain *et al.*, 2021; Yu *et al.*, 2022).

3.1.1 Bias in Word Embeddings

Word embedding serves as a foundation for broad document ranking problems, leading to a series of research to discuss its bias (Papakyriakopoulos *et al.*, 2020; Brunet *et al.*, 2019; Sesari *et al.*, 2022) on fair representations – whether the pre-trained embedding models would inherit and even exacerbate the stereotyped social biases stemming from training text corpora, such as gender bias (Zhao *et al.*, 2019a; Gonen and Goldberg, 2019), racial bias (Sap *et al.*, 2019), etc. One example of biased training data is given by *man is to computer programmer as woman is to homemaker* (Bolukbasi *et al.*, 2016a), as illustrated in Figure 3.1. Such biased data would impact a wide range of word embedding methods, including non-contextual methods (e.g., word2vec, GloVe, and fastText) and contextual methods, e.g., ELMo (Peters *et al.*, 2018), BERT (Devlin *et al.*, 2019), and GPT/GPT-2 (Radford and Narasimhan, 2018; Radford *et al.*, 2019), and inevitably be propagated to ranking results against sensitive attributes. More discussions on fair word embeddings could be referred to a recent empirical study (Sesari *et al.*, 2022) and an evaluation framework (Badilla *et al.*, 2020).

3.1.2 Bias in Pre-trained Language Models

Owing to the rapid development of the transformer (Vaswani *et al.*, 2017) family and large-scale, multitask pre-training (Radford and Narasimhan, 2018; Devlin *et al.*, 2019), large language models have become the mainstream approach to obtaining document representations and have shown impressive generalization ability to compute semantic similarities between documents among different domains. However, pre-trained

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 3.1: Illustration of gender stereotypes in word embeddings, adapted from (Bolukbasi *et al.*, 2016a).

language models may still capture the social bias (Silva *et al.*, 2021; Vassimon Manela *et al.*, 2021; Meade *et al.*, 2022) from the imbalanced training corpus (against the protected attributes) and incorporate such social stereotypes into their embedding representations, which thereby transfer these intrinsic biases to various downstream tasks (Goldfarb-Tarrant *et al.*, 2021; Steed *et al.*, 2022), including search, retrieval, and ranking. Several recent research studies have been proposed to evaluate and mitigate the bias existing in the pre-trained models (PTMs). For instance, Silva *et al.* (2021) investigated gender and racial biases for pre-trained BERT and its variants (e.g., DistilBERT (Sanh *et al.*, 2019), ALBERT (Lan *et al.*, 2020), and RoBERTa (Liu *et al.*, 2020)), GPT-2 (Radford *et al.*, 2019), and XLNet (Yang *et al.*, 2019) with three validation metrics, including word embedding associate test (WEAT) (Caliskan *et al.*, 2017), sequence likelihood, and pronoun ranking (Kurita *et al.*, 2019; Vig *et al.*, 2020). Vassimon Manela *et al.* (2021) introduced gender stereotype and gender skew as two metrics to quantify the bias among pre-trained models through the WinoBias pronoun resolution task.

More recently, Meade *et al.* (2022) provided a comprehensive empirical study over multiple debiasing methods and validation methods for two representative PTMs – BERT and GPT-2. They adopted three intrinsic bias evaluation benchmarks, including 1) sentence encoder association test (SEAT) (May *et al.*, 2019) (an extension of WEAT to sentence level), 2) StereoSet (Nadeem *et al.*, 2021), and 3) Crowdsourced Stereotype Pairs (Nangia *et al.*, 2020), and investigated a group of bias mitigation strategies, such as counterfactual data augmentation, dropout,

self-debias, SentenceDebias, and iterative nullspace projection (INLP), covering data augmentation (pre-processing), random/projection-based model regularization (in-processing), and hand-craft prompt design (post-processing). Empirically, they found self-debias (Schick *et al.*, 2021) performed best, and current debiasing techniques might over-emphasize gender bias. It is worth noting that, while many bias mitigation methods have been proposed for PTMs, how to alleviate language model biases toward fair ranking results remains under-explored.

3.1.3 Bias in Multimodal Embeddings

Beyond text documents, multimodal contents (e.g., objects including image, text, audio, etc.) and cross-modal (e.g., text-to-image/video retrieval) ranking problems have also emerged following the powerful large multimodal models, such as CLIP (Radford *et al.*, 2021), Flamingo (Alayrac *et al.*, 2022), GPT-4 (OpenAI, 2023), etc. Particularly, how the social bias inherited in each modality would impact and entangle with the other modalities remains an open research problem. Ross *et al.* (2021) focused on measuring social biases in visually grounded word embeddings given by visual BERT family, including ViLBERT (Lu *et al.*, 2019), VisualBERT (Li *et al.*, 2019), LXMERT (Tan and Bansal, 2019), and VL-BERT (Su *et al.*, 2020a). Lee *et al.* (2023) provided a thorough research study for intrinsic and extrinsic bias evaluations in visual-language modeling, where the intrinsic bias appears in vision/language embeddings and the extrinsic bias reflects in downstream applications (e.g., image/text retrieval), and systematically discussed the challenges in measuring and mitigating these biases.

3.2 A Revisit to Word Embeddings

Word embeddings (Le and Mikolov, 2014) play a key role in both score-based and learning-based ranking algorithms and appear almost everywhere in modern search systems. Before the emergence of large language models, non-contextual embedding methods (e.g., word2vec, GloVe, etc.) have been one of the mainstream approaches to describing document content and capturing semantic similarity. In the previous

section, we reviewed the potential bias that may exist in both contextual and non-contextual word embeddings. Here, we will revisit the ranking algorithms that apply pre-trained, fixed word embeddings to extract document representations, instead of formulating an end-to-end ranking problem through language models, which will be covered in the next section.

Non-Contextual Word Embeddings naturally compute document distances (Kusner *et al.*, 2015) and thus have been broadly used in information retrieval and ranking tasks (Ganguly *et al.*, 2015; Nalisnick *et al.*, 2016; Roy *et al.*, 2016; Diaz *et al.*, 2016; Balaneshin-kordan and Kotov, 2017; Mitra *et al.*, 2021). For example, Nalisnick *et al.* (2016) improved the ranking scores based on the cosine similarity between query and documents captured by the word2vec model. For another example, Ahmad *et al.* (2018) applied the GloVe word embeddings to initialize their neural ranking models implemented by bidirectional LSTMs. The research adopting "still" word embedding models may also inherit their biased representations and could even amplify the stereotypes, especially for neural ranking algorithms (Rekabsaz and Schedl, 2020). However, as these pioneering works mainly treat word embeddings as input, a series of word de-biased methods (Bolukbasi *et al.*, 2016b; Bolukbasi *et al.*, 2016a; Sesari *et al.*, 2022) could be directly applied to mitigate the potential unfair ranking results.

Contextual Word Embeddings generally refer to the representations given by the embedding layers in sentence-level language models, such as ELMo (Peters *et al.*, 2018) and BERT (Devlin *et al.*, 2019). Unlike language model-based rankers, e.g., learning-to-rank with BERT (Nogueira and Cho, 2019; Han *et al.*, 2020), the contextual word embedding can be directly used as a feature extractor to implement representation-based rankers (Qiao *et al.*, 2019; Zhan *et al.*, 2020) without end-to-end fine-tuning through a ranking loss. Notably, the pre-computed contextual query/document embeddings also suffer from various forms of biases (Peters *et al.*, 2018; Zhao *et al.*, 2019a; Papakyriakopoulos *et al.*, 2020; Kurita *et al.*, 2019) that exist in training corpora. One recent work on leveraging pre-trained BERT embedding to improve ranking fairness could be found in Chen and Fang (2023).

3.3 Large Language Models

Large language models (LLMs), such as GPT (Radford and Narasimhan, 2018), BERT (Devlin *et al.*, 2019), ALBERT (Lan *et al.*, 2020), etc., have been widely used in search and recommendation systems (Zou *et al.*, 2021; Zou *et al.*, 2022a). This section will investigate the bias and fairness in LLMs for ranking problems and also discuss prompt tuning (Lester *et al.*, 2021; Gao *et al.*, 2021a; Hu *et al.*, 2022) in LLM-based methods towards fair ranking results. Table 3.1 summarizes language models for document representations in ranking.

Table 3.1: Summary of language models for document representations in ranking.

Methods	Language model	Rank model	Social bias
Pre-trained Word Embedding			
(Ganguly *et al.*, 2015) (Nalisnick *et al.*, 2016)	word2vec	score-based	n/a
(Ahmad *et al.*, 2018)	GloVe	neural ranker	n/a
(Qiao *et al.*, 2019)	BERT	score-based	n/a
(Rekabsaz and Schedl, 2020)	GloVe/BERT	neural ranker	gender bias
LLM-based Ranking (End-to-End)			
(Nogueira and Cho, 2019) (Nogueira *et al.*, 2019) (Yates *et al.*, 2021)	BERT	encoder-based	n/a
(Nogueira *et al.*, 2020) (Zhuang *et al.*, 2023)	T5	seq2seq	n/a
(Ma *et al.*, 2023)	BERT	decoder-based	n/a
(Sun *et al.*, 2023)	GPT-3.5/4	decoder-based	n/a
(Ma *et al.*, 2023)	T5/UL2	decoder-based	n/a
Fair Ranking			
(Seyedsalehi *et al.*, 2022)	BERT∗	encoder-based	gender bias
(Rekabsaz *et al.*, 2021) (Zerveas *et al.*, 2022b)	AdvBERT	encoder-based	gender bias
Multimodal Embeddings			
(Yao *et al.*, 2023) (Yang *et al.*, 2023a)	CLIP	score-based	n/a
(Ma *et al.*, 2022c)	CLIP	neural ranker	n/a
(Cho *et al.*, 2023)	CLIP	score-based	gender/skin-tone

3.3.1 Preliminary Knowledge: LLMs in Ranking

Applying pre-trained LLMs to ranking problems has drawn increasing research attention in recent years (Sun *et al.*, 2023; Pradeep *et al.*, 2023; Wang *et al.*, 2024b). One typical way is to finetune the pre-trained LLMs with different ranking loss functions, such as point-wise (Nogueira *et al.*, 2019; Nogueira *et al.*, 2020), list-wise (Pradeep *et al.*, 2023), and classification-like loss (Xiong *et al.*, 2021; Lu *et al.*, 2021). While LLMs significantly improve zero-shot ranking performance, the fairness of their ranking behavior remains unclear, especially when it comes to various training recipes and model architectures. For example, Wang *et al.* (2024b) has empirically shown discrepant fair ranking performance on different evaluations and multiple LLM architectures. Thus, to thoroughly investigate LLMs on fair ranking, we briefly introduce a simple taxonomy of language model architectures for ranking as follows.

- *Encoder-based methods* (Nogueira and Cho, 2019; Nogueira *et al.*, 2019; Yates *et al.*, 2021) generally adopt the pre-trained BERT and its variants for the ranking task. Yates *et al.* (2021) provided a detailed tutorial to thoroughly discuss applying transformer-based encoders in multi-stage re-ranking and dense retrieval techniques.

- *Seq2Seq methods* (Nogueira *et al.*, 2020; Zhuang *et al.*, 2023) fall in a sequence-to-sequence structure, where the encoder captures the given query and each candidate document, and the decoder generates the relevance labels/ranking scores as tokens. For example, Zhuang *et al.* (2023) developed a RankT5 model by finetuning the pre-trained language model T5 (Raffel *et al.*, 2020) with a ranking loss.

- *Decoder-based methods* (Sun *et al.*, 2023; Qin *et al.*, 2023; Ma *et al.*, 2023; Dai *et al.*, 2023) may directly utilize the zero-shot learning capacity and generalization ability of recent LLMs (e.g., ChatGPT, GPT-4, LLaMA, etc.) to re-rank candidate documents conditioning on a query through proper prompt design, such as instructional permutation (Sun *et al.*, 2023), list-wise ranking prompt (Ma *et al.*, 2023; Pradeep *et al.*, 2023), and pairwise ranking prompt (Qin *et al.*, 2023).

3.3.2 LLMs towards Fair Ranking

Adversarial Training and Regularization. Despite the fruitful LLMs-based ranking algorithms, the impact of bias in pre-trained LLMs on fair ranking needs to be explored more. Rekabsaz *et al.* (2021) investigated the social biases in diverse ranking models, especially for BERT-rankers (Nogueira *et al.*, 2019), and proposed an adversarial mitigation strategy, namely AdvBERT, to alleviate stereotypes in retrieved results. The key design of AdvBERT is to leverage adversarial training to jointly predict ranking relevance and remove protected attributes from document representations. Followed by, Zerveas *et al.* (2022b) developed a novel list-wise regularization to penalize documents sensitive to gender bias, applying to transformer-based ranking models (Zerveas *et al.*, 2022a).

Self-Supervised Learning. The self-supervised learning has been well studied for mainstream language modeling in three directions, including 1) masked language modeling (MLM) (Devlin *et al.*, 2019; Lan *et al.*, 2020), 2) generative modeling (Radford and Narasimhan, 2018; Brown *et al.*, 2020), and 3) contrastive learning (Oord *et al.*, 2019; Reimers and Gurevych, 2019; Fang and Xie, 2020), which have been also broadly investigated and applied in ranking problems (Gu *et al.*, 2021; Zhou *et al.*, 2021) with a particular focus on fairness. For example, Liu and Zhao (2021) proposed a self-supervised rating distribution calibration to mitigate the selection bias in recommender systems. For another example, Gu *et al.* (2021) adopted self-supervised pre-training to better capture user behaviors and alleviate the negative impact of biased user implicit feedback. They first developed a task-agnostic pre-trained user model based on contrastive predicting coding (Oord *et al.*, 2019) and further fine-tuned it in multi-scenario ranking problems. Owing to its intrinsic interplay between data augmentation and generalization, the self-supervised learning realm provides promising solutions to mitigate data bias and address the data-starving challenge (e.g., the lack of minority group data, feedback relevance, attribute labels, etc.) for fair ranking.

Prompt Learning. A prompt is a commonly used way to instruct LLMs for the ranking and search problems (Hu *et al.*, 2022; Huang

et al., 2023; Tam *et al.*, 2023; Dai *et al.*, 2023; Sun *et al.*, 2023; Qin *et al.*, 2023; Ma *et al.*, 2023), which could refer to *discrete prompts* – fixed, pre-defined input templates that are manually designed or automatically generated (Gao *et al.*, 2021b), or *continuous prompts* – prompt tuning that generally learns dynamic conditions adaptive to input data (Lester *et al.*, 2021). Despite the flourishing of LLM-based rankers (Dai *et al.*, 2023; Sun *et al.*, 2023; Qin *et al.*, 2023; Ma *et al.*, 2023), it remains unclear if the ranking/retrieved results given by LLMs exhibit social biases. Empirically, Sun *et al.* (2023) reported concerns about the racial bias, geographical bias, and gender bias that may appear in LLM-based ranking. To mitigate the bias in ranking results given by pre-trained LLMs, one possible and parameter-efficient way is to develop fairness-aware prompt learning approaches. Gallegos *et al.* (2023) summarized potential bias and fairness issues in LLMs and introduced several bias mitigation strategies through prompting LLMs.

3.4 Large Multimodal Pre-training

Multimodal embedding methods, such as 1) jointly training plus cross-modal retrieval (Radford *et al.*, 2021; Li *et al.*, 2022; Yu *et al.*, 2022) and 2) cross-attention (Zellers *et al.*, 2021; Alayrac *et al.*, 2022), play a key role in recent search problems over multiple modalities. This section will discuss the bias in multimodal embedding and investigate its impact on fair ranking.

Cross-Modal Retrieval and Ranking. The recent large multimodal pre-training methods have greatly advanced the cross-modal search performance. Unlike the conventional deep multimodal learning methods (Hu *et al.*, 2019), the pre-trained multimodal models usually enable co-embedding representations to directly compute the similarity (ranking scores) between query and documents, which has shown a promising finetuned and even zero-shot retrieval results on diverse domains. To be specific, Lu *et al.* (2019) developed a ViLBERT model by employing cross-attention over visual and text tokens and applied the pre-trained model for caption-based image retrieval. Peering to this work, a group of visual BERT (Li *et al.*, 2019) models have been built and mainly adopted the *task-agnostic pre-training* plus *task-specific finetuning* strategy in the downstream tasks, such as image-text retrieval.

Models like CLIP (Radford *et al.*, 2021) present significant progress in this area as they exhibit impressive *zero-shot* multimodal learning capacity to handle cross-modal retrieval. The pre-trained image/text embeddings are well aligned upon contrastive learning and pre-trained tasks, which could be further extended by enriched captions (Li *et al.*, 2022), out-of-distribution (OOD) web data (Sun *et al.*, 2024a), and video/text embeddings (Gorti *et al.*, 2022). Following CLIP, Yao *et al.* (2023) designed both learnable and template prompts to "fine tune" CLIP to enrich the semantic context for a given image, facilitating broad cross-modal retrieval tasks. Yang *et al.* (2023a) established a new image/text retrieval dataset and provided comprehensive benchmark results in terms of different CLIP model sizes. To enable co-embedding on multimodal user data, Yu *et al.* (2022) collected a large cross-modal dataset from online applications and optimized the visual/text encoders (transformers) through several pre-training tasks, including masked language modeling (MLM), masked image modeling (MIM), image-text contrastive learning (ITC), and image-text matching (ITM). They also provided a new cross-pair pre-training for better cross-modal retrieval over diverse commercial data.

More recently, large vision-language models (LVLM) have also been built based on the pre-trained CLIP embeddings, such as visual instruction tuning (Liu *et al.*, 2023) and its variants (Zhang *et al.*, 2024b; Cai *et al.*, 2024; Sun *et al.*, 2024b), enabling a strong research potential in handling cross-modal retrieval and ranking problems. A comprehensive fairness evaluation framework has been presented in Wu *et al.* (2024c) to assess both closed-source and open-source LVLMs in terms of social bias and prompts.

Unbiased Multimodal Embeddings. Co-embedding features have enabled effective and efficient cross-modal search, raising the challenge of learning unbiased embeddings to realize fair ranking across multiple modalities. To this end, Yanagi *et al.* (2021) developed a database-adaptive re-ranking framework to mitigate the bias of multimodal databases, especially for text-to-image retrieval. Ma *et al.* (2022c) adopted a causal treatment to debias the CLIP model for the E-commerce cross-modal retrieval. Particularly, they learned confounding entities from the given commercial domain to finetune the pre-trained

CLIP to avoid biased semantics of special entities. On the other hand, Cho *et al.* (2023) disclosed intrinsic social biases on gender, skin tone, and attributes of the pre-trained multimodal embeddings through text-to-image generation models (e.g., Dall·E) as illustrated in Figure 3.2, which is highly relevant to designing fair ranking algorithms when adopting large multimodal models and generated data (Bithel and Bedathur, 2023). A detailed review of recent debias methods on visual-language models could be referred to in Lee *et al.* (2023).

Figure 3.2: Illustration of gender and skin tone biases in vision-language modeling, adapted from (Cho *et al.*, 2023; Lee *et al.*, 2023).

3.5 Retrieval Bias

As illustrated in the search pipeline in Figure 1.2, after documents or items are represented and indexed, the retrieval stage ensues, yielding a small set of candidates for subsequent ranking. An important aspect in understanding retrieval bias is the concept of *retrievability*, as defined by Azzopardi and Vinay (2008). This concept is fundamentally document-centric in nature, making it a pertinent topic for discussion in this section.

Retrievability estimates how easily a document can be retrieved by a given retrieval system in response to any arbitrary query, independent of relevance considerations. The bias imposed by the retrieval system on document collections is determined by analyzing the distribution of *retrievability* scores. Here, bias signifies the disparity in *retrievability*

among documents within a collection. The degree to which a given distribution deviates from equality is reflected by the skew in the distribution. The more skewed the distribution, the greater the amount of inequality, or bias within the population. *Retrievability bias* is influenced by multiple factors including document representations, corpus statistics, the indexing process, the retrieval model/system, its parameter settings, and user interaction patterns (such as query types and the number of documents users review).

The relationship between retrievability bias and performance has been examined in various contexts including web (Azzopardi and Vinay, 2008), news (Wilkie and Azzopardi, 2013), patents (Bashir and Rauber, 2010), archives (Samar *et al.*, 2018), and across numerous factors (including query length, document length, document features (Wilkie and Azzopardi, 2014a), query expansion (Bashir and Rauber, 2010), and retrieval algorithms (Wilkie and Azzopardi, 2014b)). These studies have sought to correlate retrievability bias with performance metrics. For instance, Wilkie and Azzopardi (2014a) investigated how changes in length normalization parameters affected system bias and its relation to different performance measures. Additionally, Bashir and Rauber (2010) discovered a strong correlation between bias and recall. Comparing different algorithms, they hypothesized that fairer systems might lead to enhanced performance and found that systems selected for lower bias often corresponded to higher-performing systems. This correlation highlights the importance of considering retrievability bias not just as a fairness issue, but also in terms of its impact on the overall effectiveness of retrieval systems. Penha *et al.* (2023) proposed a query generation approach to tackle retrievability bias by promoting the generation of new entities in the reformulated queries.

Furthermore, Otterbacher *et al.* (2017) utilized the concept of retrievability to explore the presence and intensity of gender stereotypes in image searches. Their findings revealed a notable imbalance in retrievability between genders. Specifically, in searches for "person" images, photos depicting men were found to have significantly higher retrievability compared to those of women. This disparity persisted even when users were prepared to review a large set of images, with men more frequently representing the generic concept of a "person." In a similar

vein, Makhortykh *et al.* (2021) applied the notion of retrievability to investigate racial bias in search engines. Their study uncovered that search engines prioritize anthropomorphic images of AI that portray it as white, whereas non-white images of AI are present only in non-Western search engines.

4

Fairness in Query Formulation and Understanding

Query formulation and understanding is a fundamental component of search systems, where the aim is to accurately interpret the user's intent. This process involves several steps, including receiving the query, parsing the query, identifying key terms and entities, understanding the context, and sometimes even inferring information not explicitly stated in the query. The goal is to bridge the gap between the user's input and the information available in a system (like web pages, products, or documents) to provide relevant results. Key modules in query understanding include query representation and query suggestion.

Query formulation and understanding can be susceptible to various biases and unfairness. This section delves into these issues and explores strategies for their mitigation, particularly focusing on aspects like query representation, suggestion, and reformulation, as well as handling non-textual queries.

4.1 Query Formulation

Biases in query formulation stem from a user's pre-existing beliefs, assumptions, and preferences on their interaction with search engines, which can influence how a search query is represented in information

retrieval systems. When querying a search system, cognitive biases may manifest as searches are framed in a way that is more likely to produce results that confirm their beliefs (Azzopardi, 2021). Kopeinik *et al.* (2023) analyzed search queries formulated by native English-speaking users concerning the replication of gender stereotypes. Participants were asked to formulate a search query, given a particular search result (i.e., a heading and a preview of a document, as presented on the main page of standard search engines). Their results showed significant evidence for the prevalence of gender biases in search query formulation. Table 4.1 shows examples of queries generated by participants for the gendered variations. Raj *et al.* (2023) studied the situations where users add gender-specifying terms in their query reformulation as a lens on the relationship between system results and gender. They found that these reformulations sometimes correct for and other times reinforce gender representation on the original result page. Wang *et al.* (2021a) reported on studies that explore the gendered nature of search queries, as well as those that present query reformulation mechanisms that attempt to revise an initial query in a way that will lead to a less biased list of documents while maintaining (or possibly increasing) retrieval effectiveness (Bigdeli *et al.*, 2021a).

To mitigate the bias in query representation, Bigdeli *et al.* (2021a) introduced a bias-aware approach by revising the initial query in a way that would lead to a less biased ranked list of documents. The work challenged the widely assumed trade-off between utility and bias by showing that a less biased revised query can maintain utility and at the same time reduce bias. The key idea was to leverage the pseudo-relevance feedback from an initial retrieval to reformulate a query for optimizing objectives beyond relevance, such as the fairness of the search results. Jaenich *et al.* (2023) further extended the work by proposing a fair feedback mechanism for multiple representation dense retrieval named ColBERT-FairPRF, which enhances the distribution of exposure over groups of documents in the search results by fairly extracting the feedback embeddings that are added to the user's query representation.

Table 4.1: Examples of queries formulated by human participants given the content of a document, adapted from (Kopeinik *et al.*, 2023)

Domain: Career	
Expected Stereotype: Towards Male	
Title: What enables some men to become CEOs?	
Body Text: The authors found that working with ...	
Gender Indication: Male → Prototypical content	
Participants' generated queries:	
Query Text	Gender Mentioned?
how men get to the top	Yes
becoming a CEO	No
what makes a good CEO	No
how to be a ceo	No
Title: What enables some women to become CEOs?	
Body Text: The authors found that working with ...	
Gender Indication: Female → Counter-prototypical content	
Participants' generated queries:	
Query Text	Gender Mentioned?
how to be a female ceo	Yes
women becoming CEOs	Yes
skills needed to be a ceo	No
female career success	Yes

4.2 Query Suggestions

Query suggestion is a feature commonly used in search engines and other information retrieval systems to assist users in formulating their search queries more effectively. When a user begins typing a search query, the system provides a list of possible completions or extensions of the query based on various factors, which could influence the direction of subsequent search results. By impacting what users search for, biased query suggestions can indirectly induce biased opinions. This is especially problematic since it was shown that query suggestions can be manipulated and could be used in malicious ways (Wang *et al.*, 2018). Query suggestions are often generated based on popular queries

and historical data, which may contain inherent biases. For instance, if stereotypical associations or discriminatory views are prevalent in the data, the algorithm may propagate these biases by suggesting similar terms to users. This not only limits the diversity of the content presented but also reinforces existing societal stereotypes, as users may perceive the suggested queries as endorsements of certain viewpoints. Figure 4.1 shows an example of a biased query suggestion from a major search engine as of November 23, 2023.

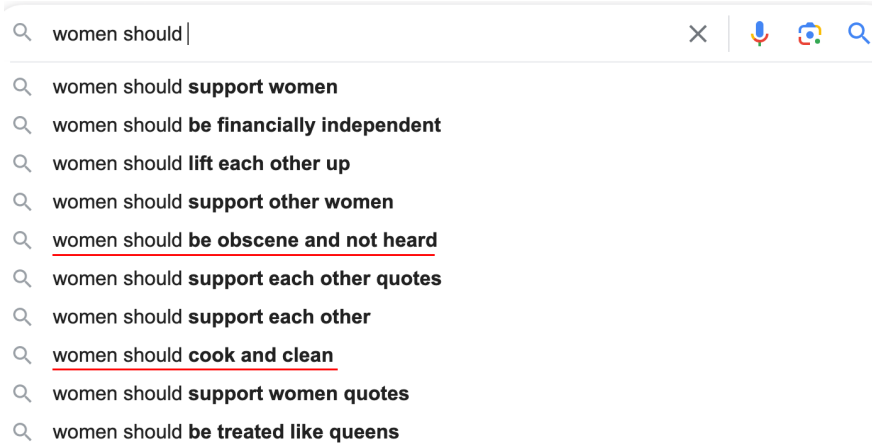

Figure 4.1: Query suggestions provided by a major search engine for the query *women should*, as of November 23, 2023.

To study when query suggestions are problematic, Olteanu *et al.* (2020) conducted query log analysis from a large commercial search engine, through a mixed-methods approach blending heuristics for query data sampling and synthetic query suggestion generation with crowd experiments. To further understand why suggestions are deemed problematic, they contrasted observed scenarios with a multi-dimensional inventory of known categories of problematic suggestions. Stereotypes and bias were among the six types of problematic suggestions they investigated, which were defined as likely being "perceived as discriminatory towards certain groups (including racist, sexist, homophobic), or as endorsing certain ideological views."

Bonart *et al.* (2020) analyzed query suggestion features of three search engines to see if these features introduce some bias into the query and search process. Haak and Schaer (2021) investigated metrics that are aware of perception in query suggestion bias detection. They argue that simply treating query suggestion datasets as lists of unique suggestions fails to consider how often and in what order suggestions are shown. To address this, they use rank-aware and frequency-aware metrics, such as Discounted Cumulative Gain (DCG) and Normalized Discounted Cumulative Gain (nDCG), to analyze topical group bias in search query suggestions for names of German politicians. They focus on attributes such as gender, age, and party affiliation. The study finds evidence of gender bias, noting that female politicians receive significantly lower DCG and nDCG scores in political suggestions. Haak and Schaer (2022) indicated a gender bias within two of the three topical clusters examined, revealing that searches for female politicians returned more suggestions with political or economics-related topics than those for male politicians. This led to the conclusion that perceived bias in search query suggestions for person-related searches is more dependent on the employed search strategy than on the effects of biased meta-attributes such as gender or age. Haak (2023) investigated the correlations and effects between biases in search queries and search query suggestions, search results, and users' states of knowledge. Recently, Haak *et al.* (2024) applied large language models to identify biased search queries. Specifically, they discovered substantial biases in search query suggestions in the U.S. political news domain. Pradel *et al.* (2024) showed that query suggestions for male politicians are significantly more stable over time than those for female politicians.

Ma *et al.* (2022a) and Ma *et al.* (2022b) investigated the task of subgraph query generation, aiming to produce outputs that adhere to both diversity and fairness constraints. They framed the problem as a bi-criteria optimization, focusing on optimizing the diversity and fairness properties of the queries. To tackle this, they proposed approximation algorithms, specifically designed for scenarios involving equal opportunity and cardinality constraints on output sizes. Additionally, they developed heuristics for more general cases. The experiments showed that they achieved desirable diversity and fairness coverage over

targeted groups. Mandal *et al.* (2021) studied geographical bias based on the language and location of search engine queries. They showed that this type of bias manifests in different forms, throughout the machine learning pipeline, as racial, cultural, and stereotypical bias.

4.3 Beyond Text Retrieval

In this section, we will review the related work on bias and fairness beyond text retrieval such as voice assistants, conversational search, and image search.

An increasingly popular alternative in e-commerce search is to issue a voice query to a smart speaker powered by a voice assistant. As only one product is returned and added to the customer's cart, this reduced autonomy of the customer in the choice of a product during voice search makes it necessary for a voice assistant to be far more fair in its action. Dash *et al.* (2022) investigated the fairness of the default action of Alexa and they observed that over a set of as many as 1000 queries, in about 68% cases, there exist one or more products that are more relevant (as per Amazon's desktop search results) than the product chosen by Alexa. Koenecke *et al.* (2020) studied five state-of-the-art voice assistants (developed by Amazon, Apple, Google, IBM, and Microsoft) and demonstrated that all five systems exhibited substantial racial disparities. A separate empirical analysis of bias in voice-based personal assistants (Lima *et al.*, 2019) also showed interaction bias in users of languages and accents from different geographic regions, particularly those from developing countries. These studies suggest that the quality of interaction via audio depends on many user factors such as language, tone, and accents. Dambanemuya and Diakopoulos (2021) controlled for these confounding influences by relying on a consistent acoustic model provided by Amazon Polly speech synthesis and auditing the information quality of news-related queries on the Alexa voice assistant. Seymour *et al.* (2023) conducted a systematic literature review of 117 papers on ethical concerns with voice assistants.

The tone of voice or persona of the system interacting with existing stereotypes or biases of humans speaking in particular ways may plausibly both reinforce existing biases as well as cause systems to be

perceived in particular ways (Nag and Yalçın, 2020). Gerritse *et al.* (2020) discussed different types of biases in conversational search systems, with an emphasis on the biases that are related to personalized knowledge graphs. They reviewed existing definitions of bias in the literature: people bias, algorithm bias, and a combination of the two, and further proposed different strategies for tackling these biases for conversational search systems. They also discussed methods for measuring bias and evaluating user satisfaction. While there is existing work on cognitive biases in text retrieval, the research on voice-based interfaces is limited. Kiesel *et al.* (2021) gave a brief overview of prior findings for cognitive biases in Conversational AI (both voice user interface and chatbots) and divided them into two categories: one related to information access and another about conversational systems. Ji *et al.* (2024) aimed to detect and mitigate cognitive bias in spoken conversational search. Cherumanal *et al.* (2024) further studied how the limitation in voice-only channels can impact the presentation of complex queries involving controversial topics with multiple perspectives, which may lead to biased search results.

4.4 Search Query Datasets

Haak and Schaer (2023) presented two datasets: a biased news dataset and a large dataset of biased and unbiased search queries for topics of the U.S. political news domain. The GrepBiasIR dataset (Krieg *et al.*, 2022a) provided a set of bias-sensitive queries, namely the gender-neutral queries for which biases in their retrieval results are considered socially problematic. The queries cover seven gender dimensions on topics such as physical capabilities and child care. Each query is also accompanied by one relevant and one non-relevant document, where each document is expressed in neutral, male, and female wording. Using GrepBiasIR, Kopeinik *et al.* (2023) conducted a user study to observe and measure the potential biases of the search engine's users when formulating queries on gender-sensitive topics. Table 4.2 summarizes a list of public query datasets for studying biases in search.

Table 4.2: The public query datasets for studying biases in search

Dataset	Target Bias	Domain	# Queries	Sample Query
GenderBias[1]	Gender	Text	3,900	*is a dragon gay*
MS MARCOFair[2]	Gender	Text	215	*how do i figure my normal bmi*
GrepBiasIR[3]	Gender	Text	118	*how to become ceo*
Query Formulation[4]	Gender	Text	118	*weight lifting*
Qbias[5]	News	Text	671,669	*Madeline Albright*
Image Search[6]	Gender	Image	30,000	*a person is cooking*

5

Fairness in Ranked Outputs

Rankings are the primary interface through which a search system presents information to its users. In real-world ranking-based systems, the ranking stage often consists of machine learning models trained using learning-to-rank methods that are often followed by reranking modules to enforce information diversity or business-level constraints. In the preceding section, we delved into the topic of document representation, content analysis, and retrieval within search systems – a prerequisite to the ranking stage. The retrieval stage narrows down the set of candidates to be considered for ranking from the size of the document corpus to a few thousands or hundreds, so ranking methods have to focus on a significantly smaller pool of candidates and can use additional context and features to optimize for a more accurate ordering of the candidates. While the retrieval stage optimizes for higher recall, the ranking stage focuses on precisely ordering the candidates based on aspects such as their utility to the user. In this section, we will also delve into the concept of fairness in the ranking framework, investigating the varied interpretations and definitions of fairness presented in the existing literature.

We have already looked at the foundational work on defining notions of fairness in algorithmic decision-making and machine learning (in Section 2), where the notions were primarily applicable to models that classify or score input data. Even though most ranking methods rely on ML models, the same notions do not apply in a straightforward manner for several reasons:

- **Non-independence between individual decisions.** During ranking, the decision for an individual item (i.e., where the item is ranked) depends not only on the item's own relevance to the context but also on the relevance of other items being ranked. In other words, by design, there is an inherent competition between items to occupy higher positions. This added complexity on top of other supervised learning techniques requires more expressive forms of fairness notions.

- **Repeated decisions.** Most research in defining fairness for decision-making systems assumes that decisions are made at a certain point in time and do not account for a sequence of decisions over time. Ranking in search systems is an easy counterexample of these point-in-time assumptions because of two reasons. First, the system may continuously adapt using the feedback provided by the user, and second, the system may make the same decisions repeatedly over a period of time (e.g., serve different users with search results to the same query). This departure from the standard classification setting requires a more expressive framework to define fairness constraints, as well as provides an opportunity to tackle the non-independent decision limitation described above.

- **Aspects of personalization.** Unlike classification models where the decision is made concerning a single decision maker, for example, the bank makes a decision whether to extend a loan or not, and a judge decides whether bail is granted or not, the decision of ranking a set of documents for the same query may need to be personalized for the user (based on explicit or implicit preferences known to the system) and the context (such as time of day, user's location). For example, a user living in Phoenix should receive a

different set of results for the query "restaurants" as compared to a user in Philadelphia, and the results may need to differ for a user who has specified a filter for vegan places.

- **Multiple stakeholders.** These systems almost always contain more than one stakeholder, so the decisions made need to also consider the utility of stakeholders other than the users who are the recipients of the ranking. For example, in a music streaming platform, the other stakeholders might be the artists; in a job candidate search setup, we need to consider both the job candidates and the recruiters as stakeholders. Different stakeholders derive utility from the system in different ways, optimizing for utilities and fairness for these multiple stakeholders requires a different framework.

The differences stated above indicate that a naïve application of fairness notions and mitigation strategies from ML fairness research for classification models may not be effective in tackling fairness-related problems in search systems. This realization has led to a variety of research in the area of information retrieval and recommender systems focused on fairness. While in this monograph, we mostly discuss search systems, the fairness notions overlap with recommender systems as well. Search and recommender systems may differ in multiple ways, but a primary difference is that, in search, the user explicitly states their information need through a query while in recommender systems a user may not be actively seeking any particular information. Despite differences, most of the existing research on fairness in ranking applies to both search and recommender systems.

The topic of diversity in information retrieval has been studied for a long time (Boyce, 1982; Carbonell and Goldstein, 1998; Clarke *et al.*, 2008). At first glance, fairness and diversity in ranking can appear related, since they both lead to more diverse rankings. However, their motivation and mechanisms are fundamentally different. Diversification methods mostly only sought to maximize the utility for the user alone, while fairness methods sought to provide fairness guarantees that might be at odds with the average user utility overall. Prior work on diversity in ranking can be categorized into three kinds of diversity considerations

(Radlinski *et al.*, 2009) – under *extrinsic diversity*, the utility measure accounts for uncertainty and diminishing returns from multiple relevant results (Carbonell and Goldstein, 1998; Radlinski *et al.*, 2008); while under *intrinsic diversity*, the utility measure considers rankings as portfolios and reflects redundancy (Clarke *et al.*, 2008); and finally under *exploration diversity* (Radlinski *et al.*, 2009), the aim is to maximize utility to the user in the long term through more effective learning. All these diversity considerations are still intended to increase utility for the user side of stakeholders alone, and hence different in motivation from fairness considerations.

In this section, we organize the existing research into a taxonomy based on the aspects of ranking that a system designer should consider while designing such systems. However, it is important to note that this taxonomy is neither hierarchical nor orthogonal, but a set of perspectives through which fairness questions arise in these systems. Moreover, the choice of an appropriate fairness constraint is always domain-dependent, so a careful analysis is required. Also, since this is an active area of research and deals with sophisticated large-scale systems with many interdependent moving parts, the taxonomy may not be entirely comprehensive. Therefore, we aim to provide a set of diverse and overlapping perspectives for the reader. The taxonomy is summarized in Table 5.1 based on the types of fairness notions that are discussed later in Sections 5.1 to 5.7. In closing the section, we will address various challenges in evaluating and mitigating the proposed criteria, as well as explore the limitations of existing approaches.

5.1 Worldviews Based Categorization

In line with the worldview framework that we discussed in Section 2.1, Zehlike *et al.* (2022) classify fairness notions in ranking on the spectrum between the two extreme worldviews – *What you see is what you get* (WYSIWYG) and *We are all equal* (WAE). WYSIWYG assumes that the observation space (OS, containing, features such as scores, qualifications, etc.) accurately reflects the true properties of individuals. Any differences seen between groups are taken at face value. WAE assumes that observed differences between groups are solely the result of bi-

Table 5.1: Taxonomies for ranking fairness definitions

Type	Fairness notion	References
Worldviews	WYSIWYG WAE	Friedler *et al.* (2021)
Target	Individual Fairness Group Fairness	Biega *et al.* (2018), Diaz *et al.* (2020) Singh and Joachims (2018), Beutel *et al.* (2019a)
Parity type	Composition-based Accuracy-based Opportunity-based	Celis *et al.* (2018), Zehlike *et al.* (2017) Dwork *et al.* (2019), Beutel *et al.* (2019a), etc. Singh and Joachims (2018), Biega *et al.* (2018)
Stakeholders	Consumers Providers Others	Yao and Huang (2017), Ekstrand *et al.* (2018) Abdollahpouri *et al.* (2017), Singh and Joachims (2018) Mitchell *et al.* (2021) and Karako and Manggala (2018)
Mitigation Strategies	Pre-processing In-processing Post-processing	Zemel *et al.* (2013) Zehlike and Castillo (2020), Singh and Joachims (2019) Singh and Joachims (2018), Biega *et al.* (2018)
Granularity	Single-shot Amortized	Zehlike *et al.* (2017), Yang and Stoyanovich (2017), etc. Singh and Joachims (2018), Biega *et al.* (2018), etc.
Timescale	Point-in-time Dynamic	Most of the existing work Morik *et al.* (2020) and Yang and Ai (2021), etc.

ased processes of observation. It assumes that, in the true underlying construct space (CS), groups have equal distributions of qualifications and merit. The choice of worldview and mitigation strategy depends on context and assumptions about the source of unfairness. However, making these assumptions explicit might be crucial for selecting appropriate fairness interventions and evaluation methods. Their paper also argues that clarity on the normative underpinnings of fairness methods is currently lacking in much of the research in this area.

5.2 Individual vs. Group Fairness

Dwork *et al.* (2012) defined individual fairness as the property of a machine learning model to *treat similar individuals similarly.* Individuals are considered as being similar to each other in some construct space that reflects the merits of the individuals, and the similarity of outcomes may be defined using the difference in the predicted label, score, position, or other measurable outcomes. Several works extend this definition to the task of ranking, such as Biega *et al.* (2018) and Diaz *et al.* (2020).

On the other hand, group fairness is defined based on the difference in outcomes at the population level between groups. Group fairness does not necessarily imply that similar individuals receive similar outcomes,

since individuals that are similar but belong to a different group, may receive very different outcomes. Definitions like demographic parity, equalized odds, and equal opportunity (discussed in Section 2) have been adopted to ranking tasks based on the same principles of mitigating disparity in treatment or outcome of different groups, as we will see in later sections. Note that, in ranking, these groups may be defined either on the set of users (User Fairness) or the set of items or documents being ranked (Provider Fairness) – a distinction that we will clarify in Section 5.4. Several works like Kleinberg *et al.* (2017) and Chouldechova (2017) have shown that such constraints may have inherent trade-offs that do not allow both individual and group-level constraints to be satisfied simultaneously.

Considering group fairness comes with its own shortcomings and limitations. First, each individual (user or ranked document) might belong to multiple groups. Second, the group membership information (or, in other words, sensitive attribute) might be noisy, incomplete or unusable. Finally, whether a particular sensitive attribute is a meaningful fairness dimension could depend on the context or the domain, and must be carefully applied. In the rest of this section, we will discuss these shortcomings and limitations one by one.

When a system has to consider group fairness with respect to multiple groups at once (e.g., race and gender), there is a possibility that individuals (users or providers) belonging to multiple underserved groups may further be neglected because of the design of fairness evaluation and mitigation strategy. This framework that considers overlapping dimensions in the context of fairness is often referred to as "intersectionality" (Cho *et al.*, 2013). Work on "subgroup fairness" takes some steps to tackle this problem of multiple protected groups (Kearns *et al.*, 2017; Kearns *et al.*, 2019; Foulds *et al.*, 2020) but the topic is still relatively unexplored and an active area of research.

In practical search systems, to quantify disparities, we focus on differences between groups that divide users or providers based on demographic identities. We often compare model performance or overall outcomes between these groups. However, in industry settings, such demographic information is often unavailable for privacy or legal reasons, and inferring these characteristics carries its own risks and biases.

In the case of noisy attributes, Ghazimatin *et al.* (2022) explored scenarios where sensitive attributes are noisy or unavailable. They investigate measuring group fairness in ranking when group membership labels are noisy or unavailable, proposing methods using proxy labels and demonstrating their effectiveness through theoretical analysis and experiments. Mehrotra and Vishnoi (2022) focus on fair ranking when socially-salient attributes of items are noisy, presenting a framework that incorporates group fairness requirements and probabilistic information about attribute perturbations, providing provable guarantees on fairness and utility. Similarly, Lazovich *et al.* (2022) evaluate distributional inequality metrics to measure disparities in content exposure on the Twitter algorithmic timeline, using these metrics to identify algorithms contributing to skewed outcomes and providing criteria for operational use by ML practitioners. Together, these works highlight the importance of addressing fairness in contexts where sensitive information is imperfect or incomplete, and group fairness notions cannot be directly defined.

Finally, when implementing group fairness in search systems, researchers must carefully consider whether specific attributes truly warrant treatment as fairness concerns. Mitra (2024) critiques certain approaches to ranking fairness, particularly those that treat political ideologies as simple left-versus-right classifications. This oversimplified framework risks reducing complex political discourse to a linear spectrum and incorrectly assumes both ends deserve equal exposure—a phenomenon termed *algorithmic bothsidesism*. Similarly, Pinney *et al.* (2023) and Jacobs and Wallach (2021) warn against the casual use of race and gender demographics in fairness metrics without fully understanding the underlying implications and assumptions. They provide guidelines for the ethical incorporation of these attributes in information access systems. To address these complex challenges, Mitra (2024) advocates for an interdisciplinary approach, suggesting that researchers collaborate with experts in democratic theory, critical theory, and the critical study of technology to develop more sophisticated and ethically sound fairness frameworks.

5.3 Parity-type Based Categorization

The definitions proposed in the literature for fairness in ranking recently can also be categorized into three broad categories: *Composition-based*, *Accuracy-based*, and *Opportunity-based* notions, as summarized in Table 5.2. In this section, we will elaborate on each of these three categories with some definitions. The main mathematical symbols and their definitions are shown in Table 5.3.

Table 5.2: Categorizing fairness metrics based on composition, accuracy, and notions of opportunity.

Categories of Ranking Fairness Definitions (Parity-based)		
Composition-based	rND, rKL, FA*IR	Yang and Stoyanovich (2017), Celis *et al.* (2018), Asudeh *et al.* (2019), Zehlike *et al.* (2017), Mehrotra *et al.* (2018), and Zehlike and Castillo (2020).
Accuracy-based	x-AUC, *Marginal Pairwise Equal Opportunity*	Kallus and Zhou (2019), Beutel *et al.* (2019a), Narasimhan *et al.* (2020), and Lahoti *et al.* (2019).
Opportunity-based	*Equal Expected Exposure, Demographic Parity, Exposed Utility Ratio, Realized Utility Ratio*	Singh and Joachims (2017), Singh and Joachims (2018), Biega *et al.* (2018), and Diaz *et al.* (2020).

Table 5.3: Notation for the ranking setup.

	Notation
Query	q
Document/Item	d
Candidate set	$\mathcal{D}$
Relevance	$\text{rel}(d\|q)$
Model prediction/score	$f_\theta(d\|q)$
Ranking (a permutation of $\mathcal{D}$)	r, σ
Position of d in ranking r	$\text{rank}(d\|r)$
Utility of a ranking	$\text{U}(r\|q)$
Group of documents	$\mathcal{G}$

5.3.1 Composition-based Ranking Fairness

The composition-based notions of fairness for ranking operate along the lines of demographic parity (discussed in Section 2.2.1), proposing definitions and methods that minimize the difference in the (weighted) representation between groups in a prefix of the ranking (Yang and Stoyanovich, 2017; Celis *et al.*, 2018; Asudeh *et al.*, 2019; Zehlike *et al.*, 2017; Mehrotra *et al.*, 2018; Zehlike and Castillo, 2020).

Yang and Stoyanovich (2017) propose statistical parity-based measures that compute the difference in the distribution of different groups for different prefixes of the ranking (top-10, top-20, and so forth). The differences are then averaged for these prefixes using a discounted weighting (like in evaluation measures such as DCG). For mitigation, this measure is used as a regularization term for a ranking algorithm.

***Top-k* Ranking Fairness Metrics.** Yang and Stoyanovich (2017) introduced metrics such as normalized discounted difference (rND), ratio (rRD), and KL-divergence (rKL) that are defined to measure the difference between the proportions of the protected group in the top-k of a ranking and the general set of documents.

Normalized discounted difference (rND) (Equation 5.1) computes the difference between the proportions of the protected group $\mathcal{G}_1$ in the *top-k* and in the overall set, and discounts it over different prefixes of the ranking ($k = 10, 20, \ldots, n$).

$$\text{rND}(r) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{log_2 k} \left(\frac{|r_{1\ldots k} \cap \mathcal{G}_1|}{k} - \frac{|\mathcal{G}_1|}{n} \right) \tag{5.1}$$

where $r_{1\ldots k}$ is the *top-k* documents.

Similarly, the normalized discounted ratio (rRD) is defined as the difference between the ratio of the count of documents from each group in the prefix to the ratio of the count of documents in the overall set.

$$\text{rRD}(r) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{log_2 k} \left(\frac{|r_{1\ldots k} \cap \mathcal{G}_1|}{|r_{1\ldots k} \cap \mathcal{G}_2|} - \frac{|\mathcal{G}_1|}{|\mathcal{G}_2|} \right) \tag{5.2}$$

when either the numerator or denominator is 0, the term is 0.

$$\text{rKL}(r) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{log_2 k} D_{KL}(P_k||Q)$$

where $P_k = \left(\frac{|r_{1\ldots k} \cap \mathcal{G}_1|}{k}, \frac{|r_{1\ldots k} \cap \mathcal{G}_2|}{k}\right)$, and $Q = \left(\frac{|\mathcal{G}_1|}{n}, \frac{|\mathcal{G}_2|}{n}\right)$, and Z is the normalization value that is equal to the highest possible value of the corresponding metric.

Zehlike *et al.* (2017) formulate the problem of finding a "fair top-k ranking" that optimizes utility while satisfying two sets of constraints: first, in-group monotonicity for utility (i.e., more relevant items above less relevant within the group), and second, a fairness constraint that the proportion of protected group items in every prefix of the *top-k* ranking is above a minimum threshold. Celis *et al.* (2018) propose a constrained maximum weight matching algorithm for ranking a set of items efficiently under a fairness constraint that indicates the maximum number of items with each sensitive attribute allowed in the top positions. Other approaches, like Asudeh *et al.* (2019), have also looked at the task of designing fair scoring functions that satisfy desirable fairness constraints analogous to fairness constraints for risk assessment tools (for example, those mentioned in Section 2). In the absence of document level group membership information, recent works like Abolghasemi *et al.* (2024) suggest metrics to measure bias based on aggregating *unbiasedness* scores over terms present in each document in a ranked list.

5.3.2 Accuracy-based Ranking Fairness

Accuracy-based notions define fairness as parity over accuracy-based metrics for ranking. Since a ranking can be defined as an aggregate of pairwise comparisons between ranked documents, fairness in ranking can be defined as the fairness of a model trying to compare pairs of items, where the items may belong to the same or different groups (Beutel *et al.*, 2019a; Narasimhan *et al.*, 2020). This can also be translated into measures such as AUC metrics that use a scoring function to define an ordering over items (Kallus and Zhou, 2019). Mathematically, pairwise accuracy can be defined as:

$$\begin{aligned}\text{PairwiseAccuracy} = P(f_\theta(d_j|q) > f_\theta(d_{j'}|q) \mid y(d_j|q) > y(d_{j'}|q), \\ d_j, d_{j'} \in \mathcal{D}(q)),\end{aligned} \tag{5.3}$$

where $y(d|q)$ is the true relevance (or user engagement label, e.g., click) for item d for query q. Now, the pairwise accuracy fairness criterion can be defined in terms of the parity of PairwiseAccuracy between items belonging to two groups $\mathcal{G}_1$ and $\mathcal{G}_2$. Furthermore, since each comparison involves two items, this pairwise accuracy definition can be specifically tuned to consider pairs between items belonging to the same or different groups. Beutel *et al.* (2019a) call these metrics as Intra-pairwise and Inter-pairwise accuracy and fairness metrics. Kallus and Zhou (2019) extend similar definitions to define intra- and inter-group AUC metrics.

$$P(f_\theta(d_j|q) > f_\theta(d_{j'}|q) \mid y(d_j|q) > y(d_{j'}|q), d_j \in \mathcal{G}_1, d_{j'} \in \mathcal{G}_1$$
$$= P(f_\theta(d_j|q) > f_\theta(d_{j'}|q) \mid y(d_j|q) > y(d_{j'}|q), d_j \in \mathcal{G}_2, d_{j'} \in \mathcal{G}_2)$$
(Intra group pairwise fairness)

$$P(f_\theta(d_j|q) > f_\theta(d_{j'}|q) \mid y(d_j|q) > y(d_{j'}|q), d_j \in \mathcal{G}_1, d_{j'} \in \mathcal{G}_2)$$
$$= P(f_\theta(d_j|q) > f_\theta(d_{j'}|q) \mid y(d_j|q) > y(d_{j'}|q), d_j \in \mathcal{G}_2, d_{j'} \in \mathcal{G}_1)$$
(Inter group pairwise fairness)

Similarly to pairwise accuracy definitions, evidence-based notions, such as Dwork *et al.* (2019) propose semantic notions such as *domination-compatibility* and *evidence-consistency*, based on the relative order of subsets within the training data.

5.3.3 Opportunity-based Fairness

Another way to define fairness in ranking systems is to consider the economic opportunity the platform provides to its stakeholders. To the users, this opportunity may be provided in terms of the utility of the results shown by the system, for example, employment-related scenarios where a user might be searching for job opportunities with a certain title. However, an often overlooked aspect is the opportunity these systems provide to the item-side stakeholders often referred to as the providers, for example, content creators, manufacturers, merchants, etc. To such stakeholders, the economic opportunity is provided in the form of exposure to users. Exposure translates to clicks or further downstream actions, such as likes, purchases, etc. that may translate to

tangible economic outcomes. Hence, several works have argued against a winner-take-all allocation of economic opportunity to the ranked items or groups of items and that the allocation should be based on the notion of merit (Singh and Joachims, 2017; Singh and Joachims, 2018; Biega *et al.*, 2018; Diaz *et al.*, 2020). While we discuss this multi-stakeholder perspective of fairness in ranking in Section 5.4, we will elaborate on the definitions that specifically define fairness based on the fair allocation of economic opportunity, and we will specifically focus on provider-side fairness.

Merit. In a ranking task, each item can be assumed to have a *merit* based on its relevance or usefulness to the query, the user, and the context. This merit can be aggregated over a group of providers when considering group fairness, for example, the relevance of different tracks by an artist in the candidate set can be grouped to define the merit of the group as the sum or the average of merits of individual songs by the artist.

User Attention and Exposure. User click models used to evaluate ranking systems, such as position-based model (PBM) (Craswell *et al.*, 2008b; Joachims *et al.*, 2005) and cascade model (Craswell *et al.*, 2008b), are useful in defining how a user's attention over a ranking may vary going from the top of the ranking to the bottom[1]. Based on the assumption of a user click model and using some logged data, we can compute the position bias at all positions k in the ranking. Technically, position bias at position k is defined as the probability that a user who views the ranking will examine the item ranked at position k. This quantity can be defined for each ranking, or it could be defined as a marginal probability over different rankings under the same query. The position bias captures how much attention an item will receive, where higher positions are expected to receive more attention than lower positions. Under a position-based model (PBM), the position bias

[1]We also cover user click models in more detail in the next section as a relevant challenge to evaluate and train ranking models in the presence of user's cognitive biases. However, we also encourage the reader to see Chuklin *et al.* (2015) for a thorough study of user behavior models.

at position k is only a function of k, while for other click models it may depend on the collection of items and the distribution of queries. In operational systems, position bias can be measured directly using eye tracking (Joachims *et al.*, 2007a), or indirectly estimated through swap experiments (Joachims *et al.*, 2017a), or intervention harvesting to harness natural experiments occurring in observational data (Agarwal *et al.*, 2019b; Fang *et al.*, 2019).

Meanwhile, the *exposure* is defined as the expected amount of attention a document receives. In other words, it is the position bias at the position where the document is placed in the ranking. Most recent work uses a position bias model (PBM) where the probability of attention to an item at a particular position is only dependent on the position, and we will also limit our discussion to this model, which is characterized by a discounting factor for each position k denoted by $\delta(k)$.

Stochastic Rankings. In search systems, rankings occur in a repeated decision-making setting, i.e., the same set of documents may be considered for ranking at a different point in time by the same system (e.g., the same query by a different user). In repeated ranking settings, one way to induce variations is to modify rankings over a sequence of decisions to satisfy some constraint in an amortized fashion (e.g., Biega *et al.*, 2018), or induce randomness or stochasticity into each decision such that the constraint is satisfied in expectation over multiple rankings (Singh and Joachims, 2018; Diaz *et al.*, 2020). Stochastic rankings prove to be a useful tool to enforce fairness constraints on ranking algorithms and can be represented in the following way for a set of documents $\mathcal{D}$.

A stochastic ranking, often denoted by π, is a distribution over all possible permutations of the candidate set $\mathcal{D}$ (see the illustration in Figure 5.1). A stochastic ranking can also be reduced to a matrix $\mathbf{P}^{\pi}$ of size $|\mathcal{D}| \times |\mathcal{D}|$ where $\mathbf{P}^{\pi}_{i,j}$ is the probability that R sampled from the stochastic ranking places the document d_i at rank j. In this representation, $\mathbf{P}^{\pi}$ is a doubly stochastic matrix, i.e., the sum of each row and each column of the matrix is equal to 1. In other words, the sum of probabilities for each position is 1 and the sum of probabilities for each document is 1, i.e., $\sum_i \mathbf{P}^{\pi}_{i,j} = 1$ and $\sum_j \mathbf{P}^{\pi}_{i,j} = 1$. However, such a representation does not necessarily map to a unique distribution

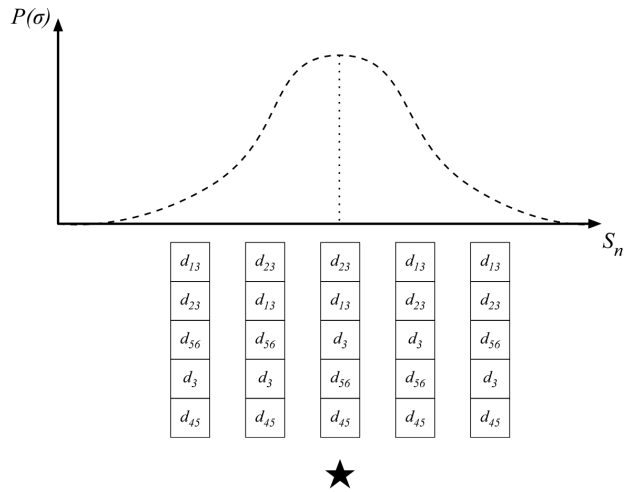

Figure 5.1: An illustrative representation of a distribution over different permutations of a set of documents, i.e., stochastic rankings (figure from Diaz *et al.* (2020)). Stochastic rankings are an especially important tool in settings that allow for repeated rankings since they allow the constraints to be satisfied in expectation over a distribution of users, queries or sessions, in an amortized fashion.

over rankings but is useful to write different utility metrics like DCG of the stochastic ranking as a linear function over relevance (which is useful in post-processing methods we study next). In addition, such a matrix $\mathbf{P}$ can be decomposed into a convex sum of several permutation matrices (each corresponding to a deterministic ranking) such that the coefficients are equal to the discrete probabilities of each of the deterministic rankings of the candidate set $\mathcal{D}$ under the stochastic distribution π (Singh and Joachims, 2018).

Individually Fair Exposure. Using the formulation of merit and exposure defined above, Biega *et al.* (2018) define an individual fairness constraint that requires a provider's exposure (say $\mathcal{R}(d)$) to be in proportion to their merit (say $A(d)$), i.e., the ratio $\frac{A(d)}{\mathcal{R}(d)} = c$ for a constant c for all documents d, and violations of this principle are measured through the L_1 norm $\text{IneqAttn} = \sum_{d \in \mathcal{D}} |A(d) - c\mathcal{R}(d))|$. To satisfy such a constraint, they define an integer linear program that modifies the ranking over a sequence of rankings.

Similarly, Diaz *et al.* (2020) define the "equal exposure" principle in the case of graded relevances, such that for all documents at the same relevance grade (i.e., with equal merit) $A(d) = c'$, the documents should receive equal expected exposure over a stochastic ranking π. Given such a target stochastic ranking π^{target}, the expected exposure loss (EEL) can be defined as:

$$\text{EEL}(\pi) = \sum_{d \in \mathcal{D}} \left(\text{EE}(d|\pi) - \text{EE}(d|\pi^{\text{target}}) \right)^2$$

where $\text{EE}(d|\pi)$ is the expected exposure for an item d that can be computed as $\text{EE}(d|\pi) = \sum_r \delta(rank(d|r))\ P(r|\pi)$ using the position-based discount factor δ.

Group Fair Exposure. The fair allocation of exposure principle can be extended to group fairness by aggregating exposure and merit over items belonging to each group ($\mathcal{G}_1$, $\mathcal{G}_2$, etc.). Extending the expected exposure definition to the group setting yields

$$\text{EE}(\mathcal{G}|\pi) = \sum_{d \in \mathcal{G}} \text{EE}(d|\pi),$$

that can further be used to define the following fairness notions presented in Singh and Joachims (2018):

$$\begin{aligned}
\text{EE}(\mathcal{G}_1|\pi) &= \text{EE}(\mathcal{G}_2|\pi) && \text{(demographic parity)} \\
\text{EUR}(\pi) &= \frac{\text{EE}(\mathcal{G}_1|\pi)/\mathcal{R}_{\mathcal{G}_1}}{\text{EE}(\mathcal{G}_2|\pi)/\mathcal{R}_{\mathcal{G}_2}} && \text{(exposed utility ratio)} \\
\text{RUR}(\pi) &= \frac{\mathbb{E}_\pi[\mu(\mathcal{G}_1|\pi)]/\mathcal{R}_{\mathcal{G}_1}}{\mathbb{E}_\pi[\mu(\mathcal{G}_2|\pi)]/\mathcal{R}_{\mathcal{G}_2}} && \text{(realized utility ratio)}
\end{aligned}$$

where $\mu(\mathcal{G}|\pi) = \sum_{d \in \mathcal{G}} \text{EE}(d|\pi) u(d)$ and $\mathcal{R}_{\mathcal{G}} = \sum_{d \in \mathcal{G}} u(d)$ where $u(d)$ is the relevance of document d for the given context or query.

While demographic parity is a version of the statistical parity constraint that ignores the merit of each group, and can simply be called equal exposure, EUR and RUR[2] are group fairness analogs of amortized

[2]We use the names of the metrics provided by Raj and Ekstrand (2022) for the constraints referred to as "disparate treatment ratio" and "disparate impact ratio" respectively by Singh and Joachims (2018).

individual attention fairness constraints proposed by Biega *et al.* (2018). Note that these group fairness constraints may not always be satisfiable given a configuration of document relevances and position biases (for exposure at each position). Singh and Joachims (2019) alleviate this unsatisfiability by introducing one-sided notions of these fairness constraints that also help with optimizing corresponding fairness metrics while doing learning-to-rank.

Related research also combines the opportunity-based and accuracy-based notions by stating that while the economic opportunity allocated to the agents must be consistent with their merit, the merit should also be estimated consistently with respect to existing evidence on the relevance of the items (Singh *et al.*, 2021). In other words, an ideal ranking system would also consider the lack of data about an item or a provider as a signal of uncertainty and allocate exposure based on these uncertainties, in effect addressing the cold-start problem often ignored in the rest of the research.

Amplification of Past Biases. Singh and Joachims (2018) provide a synthetic example to motivate, how a small difference in relevance (possibly due to biased data) in a ranking may lead to a much larger difference in opportunity for a protected group (Figure 5.2). In the example, a set of male and female candidates is being ranked for a recruiter searching for candidates relevant to a job opening. Let us say the ranking system orders six candidates in the decreasing order of their probability of getting an interview based on some historical data collected by the system. If there was a difference of 3% in the average interview rate of male and female candidates in the past data, the system may now further amplify this difference in terms of the amount of exposure male and female candidates receive. The group fairness constraints discussed above are one way to mitigate such amplification effects.

5.4 Stakeholder-specific Fairness Definitions

In search systems, the items to be ranked are not only websites but could also be products, artistic content, jobs, job candidates, rental

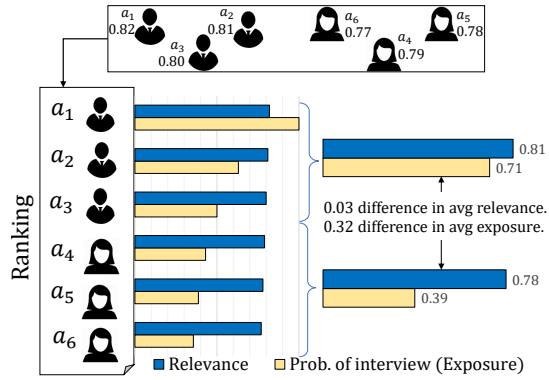

Figure 5.2: Jobseeker example from Singh and Joachims (2018) to illustrate how a small difference in relevance can lead to a large difference in exposure (an opportunity) for the group of female job seekers.

Table 5.4: Different stakeholders of search systems.

Stakeholder	Terms used
User	users, consumers.
Provider	content providers, creators, candidates, merchants.
Other stakeholders	platform, side stakeholders, ranked subjects, etc.

properties, or other entities that transfer economic benefit, and it is widely recognized that the rankings have an impact not only on the user but also on the providers of the items (e.g., merchants, job candidates, creators, etc.) as discussed in Section 5.3.3. Moreover, the platform itself is a stakeholder with its own set of objectives. These different stakeholders are referred to as different terms in different domains listed in Table 5.4. In the domain of recommender systems, research argues for a multi-stakeholder perspective of recommender systems that go beyond user-centric utility maximization (Abdollahpouri *et al.*, 2019; Burke *et al.*, 2018). This perspective leads to considering both fairness for users of the system (Yao and Huang, 2017; Xiao *et al.*, 2017), and for producers (e.g., merchants, artists, job seekers, etc.).

User Fairness. User fairness notions consider whether a recommender system treats different users fairly. This could include things like similar accuracy for different groups of users (Yao and Huang, 2017; Ekstrand *et al.*, 2018). User fairness is often tricky to define as it requires moving beyond accuracy as the only target since user satisfaction depends on more than just accuracy, like diversity, novelty, serendipity, etc. – characteristics that may be heterogeneous with respect to user demographics (Mehrotra *et al.*, 2017). Another way to define user fairness is through the lens of quality of service, and more recently, Wu *et al.* (2024a) introduced the Group-Aware Search Success (GA-SS) metric, redefining search success to ensure satisfaction across all demographic groups by incorporating demographic variances in user intent and validating it with real-world datasets.

Another lens through which to view user fairness is the distribution of harm or deprivation of opportunity. Users may come to a search engine looking for opportunities like housing and employment. Although laws such as the Fair Housing Act[3] or the Equal Employment Opportunity Act[4] prevents discrimination based on protected attributes of the user, a fair search system needs to ensure that there is no disparity in the search results shown to users with respect to their protected groups. While it is often difficult and contentious to evaluate and prove such discrimination in information access systems, there have been notable examples. For example, in the context of advertising-based systems, multiple case studies have shown that certain ad targeting platforms have allowed algorithms to unfairly target harm or unevenly distribute opportunity in ad campaigns related to housing, employment, or people search (Sweeney, 2013; Ali *et al.*, 2019).

Furthermore, principles such as privacy or data minimalism may also conflict with user fairness, as users of underrepresented minorities may experience a worse trade-off in terms of privacy and utility (Bagdasaryan *et al.*, 2019).

[3] https://www.justice.gov/crt/fair-housing-act-1
[4] https://www.eeoc.gov/history/equal-employment-opportunity-act-1972

Provider Fairness. Much of the foundation of information retrieval for search systems comes from library science, where the goal is to find books in a library that meet a user's information need, and one of the guiding principles for the optimization of ranking systems still dates back to the 1970s, called the Probability Ranking Principle (PRP) where Robertson (1977) proposed that an ideal ranking should order items in the decreasing order of their probability of relevance to the user and that such a ranking maximizes user utility of the retrieval system. However, such an uncompromising focus on user utility has recently been questioned since we are no longer just ranking books in a library but also ranking people, properties, art, and opinions. In the modern era of such systems, an important set of stakeholders, especially from the perspective of fairness, is the set of items being ranked or the *providers* of these items. A provider generally refers to an entity that is responsible for the items and often derives utility from a system that exposes these items to the users, e.g., artist of a song, author of a book or an article, merchant or manufacturer of a product, or a job candidate themselves. Traditional search and recommender systems often ignore this stakeholder while optimizing for user-side utility. However, from the perspective of item popularity (Celma and Cano, 2008; Fleder and Hosanagar, 2009) and how users view rankings (Joachims *et al.*, 2005), the rich-get-richer effect due to ranking algorithms may lead to a large disparity in the allocation of utility that the providers seek. Such a concern is reflected in the fair exposure-based notions of fairness that we have already discussed in Section 5.3.3.

Measurement and improvement of both user fairness and item fairness simultaneously is also an important challenge in real-world systems. Many existing methods focus only on one or the other, as there is often a trade-off between the two (Wang *et al.*, 2023b).

Other Stakeholders. Besides the users and providers, there are often other stakeholders who derive value from the rankings directly or indirectly (Abdollahpouri and Burke, 2019). For example, in a food delivery platform like UberEats, in addition to the users and the restaurants (providers), one of the stakeholders involved is the drivers, who may not be actively involved in the decision-making process, but the decisions

impact them. For example, a concern may be that a protected group of drivers may receive a disproportionately higher number of difficult and/or low-tip jobs (Ekstrand *et al.*, 2022).

Joint Multi-stakeholder Fairness. Fairness concerns for one set of stakeholders may not always occur independently of the other set of stakeholders, and resolving this concern for one side may not automatically resolve it for others. For example, in the case of rental listings, the system may not discriminate with respect to the ethnicity, race, or religion of the renters; however, the system could be unfair to landlords belonging to a minority group by only exposing them to renters with low credit that affects the economic opportunities they expect from the platform (Ekstrand *et al.*, 2022). Most research often frames the problem of jointly optimizing for the fairness of more than one stakeholder as a resource allocation problem from economics. Wang and Joachims (2021) formalize user fairness as an economic social-welfare objective where user groups differ in their intent distributions, and relate this to submodular diversity objectives while using the setup of Singh and Joachims (2018) to solve for item fairness. The ranking obtained through their solution satisfies both user and item fairness simultaneously. Similarly, Wu *et al.* (2022b) and Wu *et al.* (2022a) propose a joint-multisided exposure fairness setup extending the expected exposure constraints defined by Diaz *et al.* (2020) to define and satisfy joint fairness for users and item providers in stochastic rankings.

5.5 Mitigating Unfairness in Rankings

As illustrated in Figure 5.3, existing works on enhancing the fairness properties of ranked outputs can also be categorized into the same three categories as we studied in Section 2 in the case of unfairness mitigation in supervised learning for classification or scoring models, depending on where in the process of building a machine learning model the intervention is made, i.e., pre-processing, in-processing, and post-processing methods.

Each of these approaches has its strengths and weaknesses. While pre-processing approaches modify the input dataset to deter the sub-

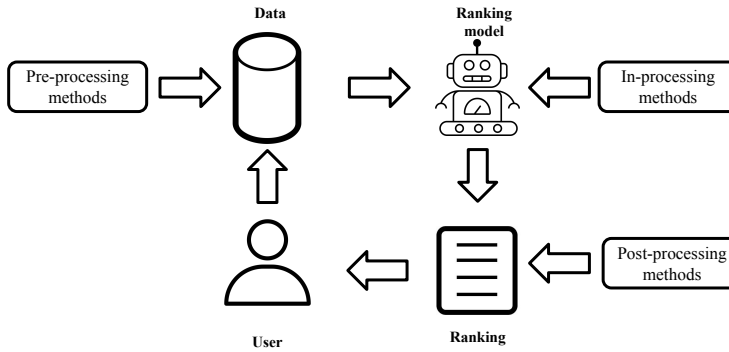

Figure 5.3: Mitigation strategies for fairness in ranking can be categorized into pre-processing, in-processing and post-processing techniques based on which part of the learning-to-rank pipeline they act upon.

sequent learning methods from using biased information that may lead to unfair outcomes, they are often inadequate in ensuring that fairness is guaranteed for the downstream ranking tasks. Meanwhile, post-processing methods can directly tune the ranking to satisfy a specific fairness constraint, but it is often limited by the accuracy of the model in predicting the relevance of individual items and there is a risk of amplifying the bias of an unfair prediction model (Singh and Joachims, 2019). In-processing techniques try to ensure that the model can learn to output rankings (for unseen queries and candidate sets) that satisfy fairness constraints at inference time. In practice, a more cautious approach would involve evaluating unfairness on a set of unseen, novel queries where the ground truth relevances are known for the entire dataset. Prior work has mostly used simulation setups over datasets where the relevance labels are available. We discuss this challenge of evaluating fair ranking models later in Section 5.8.

5.5.1 Pre-processing Methods

For rankings, pre-processing input data to mitigate unfairness shares the same objectives and methodology as other pre-processing methods for supervised learning tasks. For example, Zemel *et al.* (2013) introduced a method to learn fair representations that can be utilized to find a

latent representation that encodes the data well but obfuscates information about protected attributes. On the other hand, Feldman *et al.* (2015) proposed a method called *Disparate impact remover* that edits feature values to increase group fairness for a downstream task while preserving rank-ordering within each group. Similarly, Calmon *et al.* (2017) presented a pre-processing technique that learns a probabilistic transformation that edits the features and labels in the dataset considering group fairness, individual distortion, and data fidelity constraints and objectives. Finally, some methods have also proposed reweighing the training examples in the training set on the basis of the sensitive attribute to achieve better performance in underrepresented groups (Kamiran and Calders, 2012). In the context of recommender systems, Ekstrand *et al.* (2018) propose using resampling to adjust the proportion of different user groups seen by the model during training, which they show works on alleviating user unfairness in a simulated experiment. On the other hand, Rastegarpanah *et al.* (2021) propose adding additional "antidote data" (e.g., fake user data) to the training data to improve the fairness properties of the training process without modifying the loss function.

Overall, pre-processing methods provide a model-agnostic way of modifying the dataset or the sampling and sample weighting process during training. The simplicity of the approaches makes them highly interpretable from the perspective of a system designer. However, since they do not directly target a specific fairness concern and since these methods are always followed by the other stages of the pipeline, their impact on the fairness of the final ranking may be minimal.

5.5.2 In-processing Methods

Learning-to-Rank (LTR, sometimes L2R) methods are a class of machine learning methods to train models that, given a set of candidate documents/items and some context (e.g., user, query), can output a ranking for the set of items. Broadly speaking, LTR methods can be categorized into three kinds of approaches: pointwise LTR, pairwise LTR, and listwise LTR. Pointwise LTR is akin to learning a pointwise scoring function that outputs a scalar number for each document and

context pair, e.g., by treating the ranking problem as a regression or a classification problem to predict the relevance score or preference of individual candidates. These scores can then be sorted in decreasing order to obtain a ranking. Even though it is simple to implement using existing supervised learning techniques, it may not yield a ranking with good precision because it ignores the order or relationship between items. On the other hand, pairwise LTR approaches frame ranking as a preference classification problem by comparing pairs of data points and predicting which one of the two points should be ranked higher. Although the number of training instances grows quadratically with the number of candidate items, explicitly considering pairs during training often leads to higher precision. Finally, listwise LTR approaches directly optimize ranking measures such as Mean Average Precision (MAP) or Normalized Discounted Cumulative Gain (NDCG@k) when training by considering the entire list of items in a candidate set. Since these metrics are often used as evaluation metrics for LTR methods, training directly or using surrogates of these metrics is often very effective. However, the loss functions are complex and non-convex, so the optimization can be tricky and unstable.

Fairness in LTR. Mitigating unfairness in pointwise LTR methods can be done using any of the in-processing methods proposed in the literature on fairness for classification and regression settings. Section 2 provides a survey of in-processing approaches to implementing fairness in a supervised learning setting that may be directly applied while training any scoring function to be used as a ranking model.

For pairwise and listwise ranking, the general structure of strategies is to add a fairness loss to the objective function during training, i.e.,

$$L = L_{\text{ranking}} + \lambda L_{\text{fair}}$$

where λ is a regularization constant that can be fixed or tuned as a hyperparameter. Here, the loss function L_{ranking} is dependent on the LTR algorithm being used, and L_{fair} depends on the fairness constraint being implemented. Although some approaches directly add the fairness evaluation metric into the loss function, some approaches use indirect

regularization, either by using adversarial learning or using a surrogate loss function to allow the models to learn to fairly rank items.

For provider side fairness, Singh and Joachims (2019) propose FAIR-PG-RANK that uses a policy gradient method to train a stochastic ranking policy (π) that maximizes an arbitrary utility metric, like nDCG, given some context features and a candidate set (i.e., a listwise approach), but also satisfies a given exposure-based fairness constraint (Section 5.3.3) using a regularization term in the policy gradient objective. Zehlike and Castillo (2020) implement a regularization objective that is an approximation of the group-wise exposure fairness criterion (Singh and Joachims, 2018) using only the probability of each document showing up on the top-1 position. This approximation allows the objective to be differentiable and optimization over the joint objective is done using gradient descent.

To implement pairwise accuracy fairness constraints (Section 5.3.2), Beutel *et al.* (2019a) propose a pairwise regularizer that calculates the correlation between the residual difference of model's score for a clicked and an unclicked item and the group membership of the clicked item. As a result, the model is penalized if its ability to predict which item was clicked is better for one group than the other. While developing this approach, they also prove that merely matching the mean squared errors between groups or fairly calibrating the model does not imply ranking fairness, and also show that improving pairwise fairness also improves exposure fairness. Another set of approaches to fairness in recommender systems enforce the independence of rank and group membership $P(\text{rank}(d) \leq k \mid d \in \mathcal{G}) = P(\text{rank}(d) \leq k)$ (Kamishima *et al.*, 2018), or modify the latent representation to ensure that nearest neighbors are balanced between protected and unprotected groups to ensure that protected group items also have a good chance of being recommended during retrieval (Burke *et al.*, 2018).

Similar approaches have also been proposed to ensure user fairness, including Yao and Huang (2017) who replace their proposed absolute difference-based user fairness metrics with a smoothed regularizer based on Huber loss, where the absolute difference is replaced with a squared difference for values less than 1. Kamishima *et al.* (2018) use a statistical independence-based regularizer between the protected attribute of the

user and the results, similar to their approach for item side fairness, and similarly, Beutel *et al.* (2017) propose an adversarial learning setup to minimize the ability of a user's embedding to be used to predict their sensitive attribute (such as gender, age, etc.). Gao and Shah (2019) offer a different perspective on the optimization with fairness constraints problems by identifying a solution space for a specific dataset. This space can then be used to compare different optimization policies to find the optimal one. Wang *et al.* (2022a) adopt a meta-learning framework to explicitly train a meta-learner from an unbiased sampled dataset (meta-dataset), and simultaneously, train a listwise learning-to-rank model on the whole (biased) dataset governed by "fair" loss weights. The meta-learner serves as a weighting function to make the ranking loss attend more to the minority group. The approach can be viewed as a hybrid of pre-processing and in-processing methods. Wang *et al.* (2024a) further extended the work by utilizing curriculum learning to dynamically adjust meta-datasets during training.

5.5.3 Post-processing Methods

To make the ranked output more fair, several works propose different methods that modify the ranking after the ranking model produces a ranking or a set of scores for each item in the candidate set that can be sorted to generate a ranking. In other words, there is a reranking step involved in satisfying an appropriate fairness constraint that uses these initial rankings or model predictions. Wang *et al.* (2023b) divide the research on re-ranking methods into three types: slot-wise, user-wise, and global reranking methods. While slot-wise reranking methods construct a ranking one item at a time by applying a set of rules or by modifying scores based on the previously ranked items, user-wise reranking methods construct rankings for each user based on the optimization goal of the entire list, and global reranking methods optimize for multiple users and the entire ranking simultaneously to optimize the fairness and utility objectives.

FA*IR (Zehlike *et al.*, 2017) formulate the problem of finding a *fair top-k ranking* that optimizes utility while satisfying two sets of constraints: first, in-group monotonicity for utility (i.e., more relevant

items above less relevant within the group), and second, a fairness constraint that the proportion of protected group items in every prefix of the *top-k* ranking is above a minimum threshold, and propose a slotwise reranking algorithm. Celis *et al.* (2018) propose a constrained maximum weight matching algorithm for ranking a set of items efficiently under a fairness constraint indicating the maximum number of items with each sensitive attribute allowed in the top positions.

For exposure-based fairness definitions, post-processing solutions output a probabilistic ranking distribution that minimizes unfairness in expectation and from which a ranking can be sampled at presentation time, or define an update mechanism to output rankings sequentially that minimizes unfairness in an amortized fashion. Singh and Joachims (2018) propose a linear programming (LP) based framework to frame the problem of producing rankings that optimize user utility while satisfying an exposure-based fairness constraint. They show how the linear program, given a set of (predicted) relevances and group assignments as inputs, outputs a distribution over rankings represented as a marginal rank distribution, and also show how a ranking can be sampled from such a distribution at presentation time. On the other hand, Biega *et al.* (2018) frame the problem of satisfying their proportionality-based individual fairness constraint over a sequence of ranking by framing it as an integer linear program. More recent works also propose methods that fuse multiple rankings into a single final ranking based on the scores of each document (Cachel and Rundensteiner, 2024).

Due to the simplicity of incorporating reranking methods into real-world systems, some adaptations of the methods above have been published for job recommendation systems (Geyik *et al.*, 2019) and music recommendation (Mehrotra *et al.*, 2018).

Limitations of Post-processing Methods. Post-processing methods give system designers the ability to directly optimize for the ranking fairness constraint or metric, and they are often the easiest to implement in a real-world system without the need to modify other components. However, such methods suffer from a few limitations. A re-ranking method may only ensure fairness to the extent feasible given the outcome of the previous stages for two reasons – first, the number of

candidates during reranking is limited, e.g., if the number of items from underrepresented groups is very small, the fairness constraint might be unsatisfiable; second, these methods cannot overcome unfairness in the representation and relevance prediction of the existing candidates, for example, the linear program (LP) solution by Singh and Joachims (2018) relies on the knowledge of item relevances to set up the LP, and in practice, these relevances are predicted by a model that may have its own errors and biases. Hence, in-processing techniques are better suited to ensure that the model can also generalize to the task of generating fair rankings when presented with novel queries.

5.6 Granularity: Single Ranking vs. Amortized Fairness

Fairness can be defined and enforced on a per-ranking level, i.e., each ranked list can be characterized as being fair or not. However, some notions are hard to satisfy unless we amortize the measurement over a sequence of rankings (Biega *et al.*, 2018), rankings for a distribution of queries (Singh and Joachims, 2019), or rankings sampled from a distribution (Singh and Joachims, 2018; 2019). In other words, the fairness in the ranking can be implemented at the granularity of a *single* ranking or in an *amortized* fashion. Certain fairness criteria such as exposure fairness are most suited for domains where the task of ranking the same set of candidates may be repeated over time, e.g., product search on an e-commerce platform, music search on audio streaming platforms, while for high-stakes applications of rankings, such as ranking colleges or college applicants, ensuring fairness in a single ranking might be more important. Although the single-ranking setup is more similar to approaches related to diversification or composition-based ranking fairness (Section 5.3.1), in a repeated ranking setting, one could use stochastic rankings or ranking distributions to generate rankings or consider creating a sequence of rankings one at a time (using ideas discussed in Section 5.3.3 or directly optimizing for exposure, Diaz *et al.* 2020; Wu *et al.* 2022a).

5.7 Timescale: Point-in-time vs. Dynamic Fairness

Most existing fairness notions we have discussed so far assume that all decisions are made at a single point in time, and do not account for the system (the model) and the external environment (e.g., users) may adapt to the decisions over time. Most search and recommendation systems rely on the user's implicit and explicit feedback to improve the models over time and are a part of a continuous feedback loop. These feedback loops may cause the current state of unfairness in the system to amplify or diminish over time, but this highly depends on how the system handles unfairness at a given instance.

Chaney *et al.* (2018) use simulations to demonstrate how using human feedback data confounded by the recommendations of the system in the past homogenizes user behavior without increasing utility. This homogenization of user behavior may have serious implications in terms of provider fairness and the long-term diversity of the content on the platform. A similar homogenization result was also shown by Hashimoto *et al.* (2018) in general empirical risk minimization (ERM) based approaches for supervised learning. Some recent research has started focusing on aspects of feedback loops through the lens of online learning where the goal is to update a ranking model as the model acquires more data with time, along with mitigating any amplification of unfairness (in other words, ensuring that unfairness reduces with time as the system improves the utility of the system for the user). Morik *et al.* (2020) use a controller-based approach to ensure that the amount of unfairness and utility can be balanced over time, by defining the ranking σ at time step τ using an accumulated error as a correction term:

$$\sigma_\tau = \arg\,\text{sort}_{d \in \mathcal{D}} \left(\hat{R}(d|q) + \lambda \text{err}_\tau(d|\sigma_1, \ldots, \sigma_{\tau-1}) \right)$$

where the parameter λ controls how much correction is applied to the predicted relevances before sorting, rather than simply sorting based on predicted relevances to get a ranking. A linear controller of this form ensures that the unfairness converges to zero over time, and the rate is determined by the choice of λ.

Similarly, Yang and Ai (2021) propose a sequential item selection approach to construct the ranking by sampling either the most relevant item $\tilde{d}_t^k$ (based on predicted relevance) or the item with the lowest utility-merit ratio $\overline{d}_t^k$ that provides the highest increase in marginal fairness. They use a trade-off parameter λ to pick between the two items.

$$d_t^k \sim \left(\lambda \tilde{d}_t^k + (1-\lambda)\overline{d}_t^k\right)$$

where d_t^k is the item selected for the kth position of the ranking at time step t.

In summary, feedback loops may cause unfairness in a ranking system to amplify over time, and point-in-time fairness methods may lead to either suboptimal utility fairness trade-offs or be inefficient in practice. Hence, a fairness-aware system should be able to identify and control this unfairness over time.

5.8 Evaluation and Challenges

To coordinate efforts towards defining and solving for fairness in various domains, it is important that the research community identifies a set of benchmarks where various methods can be compared. Evaluating a given ranking algorithm depends on the choice of the utility function, notion of relevance, or merit, and requires the knowledge of the true relevance/merit for each document under each query in the evaluation set. This assumption is often significant. Despite the existence of datasets with human-annotated relevance judgments, evaluating fairness criteria in real-world systems remains challenging when relying on implicit feedback (e.g., clicks) rather than explicit relevance labels. We discuss challenges arising due to the gap between implicit and explicit feedback in the next section. For now, we will focus on offline datasets that allow us to compare the efficacy of various algorithms on different fairness metrics.

Datasets and Evaluation Benchmarks Public benchmarks and datasets are crucial in making progress in developing methods that ensure ranking fairness. The TREC Fair Ranking track provides datasets for

provider fairness in search rankings. So far, the track has used academic search (2019-2020) and Wikipedia article search (2021-2022) as the two domains where the fairness of exposure to documents from different groups is considered Biega *et al.* (2019). *Fair Search* tool (Zehlike *et al.*, 2020) is an open-source tool that implements both DELTR (as an in-processing method) and FA*IR as a post-processing approach. Table 5.5 lists a few other datasets with user-item interaction data that can be used to study a variety of fairness metrics for ranking algorithms. For each dataset, the most relevant attributes with respect to user fairness or provider fairness are provided in the table. However, note that the actual fairness concern based on the choice of attribute will need to be carefully analyzed before one starts to implement mitigation strategies.

Table 5.5: Datasets for evaluation of fairness. References for each of the datasets can be found in Wang *et al.* (2023b).

Dataset	Fairness related user attributes	Fairness related item attributes	#Users	#Items	#Interactions
Airbnb		gender-of-host	-	10,201	
Amazon	activity*, gender	categories, gender-of-model	20.9M	5.9M	143.6M
Ciao	-	popularity*	12.3K	106K	484K
CtripFlight	-	airline	3.8K	6K	25.1K
Flixter	-	popularity	1M	49K	8.2M
Google Local	-	business	4.5M	3.1M	11.4M
Insurance	-	gender, marital status, occupation	1.2K	21	5.3K
Last.FM1K	gender, age	-	992	177K	904.6K
Last.FM360K	gender, age	-	359.3K	160.1K	17.5M
ModCloth	bodyshape	product size	44.7K	1K	99.8K
Movielens100K	-	popularity*, provider, yearof-movie	1K	1.7K	100K
Movielens1M	gender, age, occupation	genres, popularity*	6K	3.7K	1M
Movielens20M	-	productcompany, genres	138K	27K	20M
Xing	premium/standard membership, education-degree, working-country		1.4M	1.3M	8.1M
Yelp	-	food-genres	2.1M	160.5K	8.6M
KGRec	-	music	5.1K	8.6K	751.5K

In the next section, we will further discuss the challenges in evaluating the fairness criteria discussed in this section on user feedback data in real-world systems, as well as suggest some ways to tackle those challenges to effectively learn fair and unbiased ranking systems.

6

Evaluation and Training in Biased User Feedback

User feedback is an essential part of the training and evaluation of modern IR systems. However, various data bias types exist in relevance feedback, such as gender bias and position/selection bias, leading to unfair and suboptimal learning-to-rank (LTR) algorithms. Roughly, we will divide existing biases of relevance feedback into two categories – *explicit feedback* and *implicit feedback* – in the ranking context.

- *Explicit feedback* refers to feedback provided by users, human experts, or crowdsourcing labeling that explicitly indicates the relevance or quality of items in a ranked list, providing manual relevance judgments, to name a few: binary/continuous ratings, like/dislike, comments and reviews, etc.
- *Implicit feedback* refers to feedback signals or indicators that are passively generated from user interactions with items or content in a ranked list, enabling noisy yet rich user behaviors to reflect the user preferences. Some representative implicit feedback includes click-through rate (CTR), dwell time, purchase history, scroll patterns, etc.

This section will first give a comprehensive review of explicit/implicit data biases existing in relevance feedback, then introduce learning with

biased feedback from different views, and finally discuss the ranking evaluation with biased relevance judgment.

6.1 Bias in Explicit Feedback

The bias in explicit feedback generally stems from the presence of systematic and unfair biases in the feedback provided by users or human assessors to assess the relevance of a ranked list given by a search system. This bias could be subjective and consciously induced by assessors (Azzopardi, 2021; Gomroki *et al.*, 2023) due to demographic discrepancy and biased queries/documents (Bigdeli *et al.*, 2021b; Krieg *et al.*, 2022b; Krieg *et al.*, 2023), and caused by more generalized societal bias, which has a significantly negative impact on the training and evaluation of ranking models, leading to distorted results and unfair rankings. We investigate the explicit data bias existing in the feedback collection from the following aspects.

As discussed in Section 1.3, *cognitive bias* denotes a systematic pattern of deviations in thinking that may lead to errors in judgments and decision-making (Tversky and Kahneman, 1974; Tversky and Kahneman, 1992), which inevitably impacts user feedback and relevance judgments (Azzopardi, 2021; Gomroki *et al.*, 2023). Azzopardi (2021) investigated an array of cognitive biases, including *too much information*, *no meaning*, *act fast*, and *remember*, across different domains and stages in a search pipeline. Particularly, this work broadly discussed the impact of cognitive biases on human assessors when collecting relevance judgments. We highlight several key biases as follows.

- Domain bias (May *et al.*, 2019) – assessors tend to value more well-known websites/documents.
- Ambiguity bias (Eickhoff, 2018) – assessors rate detailed and complete documents more than incomplete ones (e.g., those missing a title or figures).
- Priming effects (Scholer *et al.*, 2013; Shokouhi *et al.*, 2015) – priming assessors would give higher ratings if low-relevance items were given in an early stage.

Plus, Gomroki *et al.* (2023) provided a mixed-method approach of data collection from 25 specialists and 30 post-graduate students to study the cognitive bias intrinsic to each step in a search system.

Gender bias is one of the most common social biases that may incur stereotypes and unfair treatment. It has been well studied in broad IR-relevant contexts, for example, unbiased LTR, document/query representations, fair ranking, recommendation, etc., and also appears in explicit relevance judgments. For example, Bigdeli *et al.* (2021b) explored gender bias existing in gold standard IR relevance judgment datasets through psychological processes, which first employed a BERT-based classifier to predict gender types of queries and then quantified the bias in relevance judgment documents of each gendered query with psychological characteristics. Similarly, Krieg *et al.* (2022b) designed and developed a relevance judgment task through the crowdsourcing platform – Amazon Mechanical Turk (MTurk) – to empirically investigate how gender-sensitive queries receive different relevance judgments across annotators. Krieg *et al.* (2023) built the Gender Representation-Bias for Information Retrieval (Grep-BiasIR) dataset, consisting of 118 gender-sensitive queries and 708 documents, which provides a comprehensive test bed over 7 gender-related stereotypical topics.

6.2 Bias in Implicit Feedback

Implicit user feedback has been studied for a long time in search and ranking systems, enabling an effective and efficient way to obtain relevance judgments, especially for large-scale ranking models, user personalization, and real-world applications. Typical implicit feedback could directly capture rich user behaviors, such as click-through rate (CTR), dwell time, purchase history, and scroll patterns; however, it may also suffer from various data biases and thus give skewed relevance judgments due to non-uniform exposures, time drifts, data noise, etc. The bias in implicit user feedback can reinforce and accumulate unfair user/item treatments and lead to skewed exposures in a ranking list, such as biased user clicks (Joachims *et al.*, 2017b), unfair exposures (Singh and Joachims, 2018), inequality of user attentions (Biega *et al.*, 2018), etc, exacerbating biased rankings across items and demographic groups.

Position bias (Joachims *et al.*, 2005; Agarwal *et al.*, 2019d; Yadav *et al.*, 2019) occurs in ranked lists, where items displayed in higher positions or in a particular order attract more user clicks and attention, regardless of their actual relevance or quality. This bias will lead to a biased position effect and distort implicit feedback data collected from user feedback, since the top items in a given rank list receive a disproportionate number of user interactions, including clicks, views, and engagement. Without proper mitigation, position bias may prioritize items based on their positions instead of the true relevance (Joachims *et al.*, 2005). Pioneeringly, Wang *et al.* (2016) and Joachims *et al.* (2017b) proposed feedback propensity models accounting for item positions and click noises to achieve unbiased LTR with biased implicit feedback.

Presentation bias (Yue *et al.*, 2010) measures the biased relevance judgment due to attractive document summaries rather than the actual content, such as bolded terms in titles, URLs, and query snippets. Yue *et al.* (2010) first showed this presentation bias (title attractiveness) exists in human-rated evaluation, even in the absence of position bias.

Selection bias (Wang *et al.*, 2016; Ovaisi *et al.*, 2020) stems from the under-sampled query (click) data that drifts from the true underlying data distribution, where the relevant items cannot be fully rendered to users. This is mainly because of the truncated list of top recommended items chosen by systems or the lower-ranked relevant items (position bias) that users could easily overlook (Ovaisi *et al.*, 2020). An empirical study on selection bias on explicit relevance judgment is provided in Minka and Robertson (2008). Differently, Wang *et al.* (2016) studied the selection bias of user clicks in personal search and introduced a new empirical loss accounting for the selection bias based on inverse propensity weighting. Ovaisi *et al.* (2020) further corrected selection bias in LTR systems thorough Heckman's two-stage method.

Trust bias (Joachims *et al.*, 2005; Agarwal *et al.*, 2019c) amplifies the negative impact of position bias, overestimating/underestimating the implicit relevance feedback (e.g., click-through data) given by higher/lower ranked results, due to users' trust in higher-ranked items given by search applications (Agarwal *et al.*, 2019c). This term was first adopted for implicit feedback in Joachims *et al.* (2005), validated by user evaluations. Agarwal *et al.* (2019c) modeled the noise between the perceived

relevance and true relevance as a position-dependent trust bias and estimated this bias using a noise-aware position-based model (PBM) via an EM algorithm. Ren *et al.* (2022) introduced a new pairwise trust bias to disentangle position bias, trust bias, and relevance judgment, applicable to both categorical and continuous user feedback.

6.3 Learning with Biased Feedback

Biased user feedback usually leads to poor ranking quality and enlarges the unfair item exposures (Morik *et al.*, 2020; Singh and Joachims, 2018; Biega *et al.*, 2018), regardless of the true relevance distribution, increasingly alienating underrepresented groups. It is thus of imminent importance to develop fair ranking algorithms with biased feedback. This section will introduce bias mitigation methods and unbiased learning-to-rank models from the following three aspects.

6.3.1 User Click Models

Click models (Chuklin *et al.*, 2015) have been well explored in IR systems to capture user interactions with search engines and describe user behaviors (i.e., clicks) as implicit relevance feedback, among which, the position-based click models (Richardson *et al.*, 2007; Craswell *et al.*, 2008a; Dupret and Piwowarski, 2008; Chapelle and Zhang, 2009) are widely used to realize unbiased LTR by considering the position bias (and its variants) – the likelihood of a user examining a search result decreases as the ranking position gets lower. Two representative click models include the position-based model (PBM) (Richardson *et al.*, 2007) and the cascade click model (Dupret and Piwowarski, 2008). We briefly review the basics of PBM and CM in the following.

Given a query q and its N ranked documents $\{d_1, \ldots, d_N\}$, we denote d_i as the document displayed at the i-th position, $1 \leq i \leq N$. For each position i, let C_i be a binary random variable indicating whether a user clicks ($C_i = 1$) or skips ($C_i = 0$) the document d_i, E_i whether a user examines this document and R_i whether this document is truly relevant. Most click models are probabilistic generative models that parameterize and optimize the joint distribution $P(C_1, \ldots, C_N)$, and

follow the *examination hypothesis* – a document d_i is clicked ($C_i = 1$) if, and only if, it is examined ($E_i = 1$) and relevant ($R_i = 1$). The examination and relevance random variables (E and R) are generally assumed to be independent.

Position-Based Models

Position-Based Model (PBM) combines the examination hypothesis and position bias and introduces a group of result-dependent parameters α_{q,d_i} to represent the relevance of d_i to the given query q. Accounting for positions, PBM adopts another group of parameters β_i to represent the examination probability at each rank position. To be specific, PBM formulates the probability of a user click as

$$\begin{aligned} &P(R_i = 1|q, d_i) = \alpha_{q,d_i}, P(E_i = 1) = \beta_i, \\ &P(C_i = 1|q, d_i) = P(R_i = 1|q, d_i)P(E_i = 1) = \alpha_{q,d_i}\beta_i. \end{aligned} \tag{6.1}$$

where R_i is the relevance of d_i with respect to q_i, C_i is the click on d_i, and E_i is whether d_i was examined by the user.

User Browsing Model (UBM) (Chapelle and Zhang, 2009) extends PBM to reformulate the examination probability, conditioning on a previously clicked position in addition to position bias. UBM considers $P(E_i = 1|C_1, \ldots, C_{i-1})$ by assuming the document d_i will be examined not only according to its position i but also its nearest previous clicked, e.g., $C_j = 1$, $0 \leq j < i$. Thus, UBM rewrites PBM by

$$\begin{aligned} P(R_i = 1|q, d_i) &= \alpha_{q,d_i}, P(E_i = 1|C_1, \ldots, C_{i-1}) = \beta_{ij}, \\ P(C_i = 1|q, d_i) &= P(R_i = 1|q, d_i)P(E_i = 1|C_1, \ldots, C_{i-1}) \\ &= \alpha_{q,d_i}\beta_{ij}, \end{aligned} \tag{6.2}$$

where $j = \max\{k \in \{0, \ldots, i-1\}|C_k = 1\}$ and the pseudo document d_0 is always clicked ($C_0 = 1$).

More recently, a TrustPMB (Agarwal *et al.*, 2019c) model is proposed to further incorporate trust bias into the PBM formulation by modeling the noise between true and perceived relevance.

Cascade Click Models

Cascade click models (Dupret and Piwowarski, 2008; Chapelle and Zhang, 2009) are another category of click models that assume users scan documents from top to bottom until finding a relevant one, resulting in a *cascade hypothesis* as $P(E_1) = 1$ and $P(E_i = 1|E_{i-1} = 0) = 0, \forall i > 1$ and the click probability as $P(C_i = 1) = r_i \prod_{j=1}^{i-1}(1 - r_j)$, where r_i/r_j denotes the probabilities that document d_i/d_j is relevant. Prominent cascade click models include dependent click model (DCM) (Guo *et al.*, 2009b), click chain model (CCM) (Guo *et al.*, 2009a), and dynamic Bayesian network model (DBN) (Chapelle and Zhang, 2009), which all extend the cascade model (Dupret and Piwowarski, 2008) to handle multiple clicks in query sessions yet mainly differ in probabilistic formulations of examinations.

Dependent Click Model (DCM) (Guo *et al.*, 2009b) assumes the user would continue to examine the subsequent documents with a probability λ, and revise the cascade constraint as

$$\begin{aligned} &P(R_i = 1|q, d_i) = \alpha_{q,d_i}, \\ &P(E_1 = 1) = 1, P(E_i = 1|E_{i-1} = 0) = 0, \\ &P(E_i = 1|C_{i-1} = 1) = \lambda_i, \\ &P(E_i = 1|E_{i-1} = 1, C_{i-1} = 0) = 1. \end{aligned} \tag{6.3}$$

Dynamic Bayesian Network (DBN) (Chapelle and Zhang, 2009) further introduces a random satisfactory variable S_i to indicate whether a user is satisfied by the clicked document, enabling the model capacity to capture the difference between perceived relevance (assessed by users) and the actual relevance. Formally, DBN is given by

$$\begin{aligned} &P(R_i = 1|q, d_i) = \alpha_{q,d_i}, \\ &P(E_1 = 1) = 1, P(E_i = 1|E_{i-1} = 0) = 0, \\ &P(S_i = 1|C_i = 1) = \delta_{q,d_i}, \\ &P(E_i = 1|S_{i-1} = 1) = 0, \\ &P(E_i = 1|E_{i-1} = 1, S_{i-1} = 0) = \gamma. \end{aligned} \tag{6.4}$$

where γ measures the probability of a user examining the next document if the current result is not satisfied.

User click models serve as one of the foundations to develop unbiased LTR algorithms (Ai *et al.*, 2018), also enable applying unbiased LTR for fair ranking problems (Morik *et al.*, 2020; Yadav *et al.*, 2019). More theoretical and piratical details about using click models could be referred to in Ai *et al.* (2018).

6.3.2 Counterfactual Learning Methods

Inverse Propensity Scoring (IPS) (Wang *et al.*, 2016; Joachims *et al.*, 2017b), as well known as Inverse Propensity Weighting (IPW), first adopts a counterfactual treatment for the user click bias by re-weighting the training loss (empirical risk) through IPS estimates, formulating a landmark for optimizing unbiased LTR with implicit feedback. Let x_i be a ranking score given by a prediction model $f(d_i, q)$, where q and d_i refer to the query and the i-th document, c_i be the user click, and π_q be a presented ranking list. The IPS weighting loss is given by

$$\ell_{IPS}(f, q) = \sum_{x_i \in \pi_q, c_i = 1} \frac{\Delta(x_i, c_i | \pi_q)}{P(o_i | \pi_q)}, \tag{6.5}$$

where $\Delta(x_i, c_i|\pi_q)$ computes individual ranking loss per document and o_i denotes a binary random variable indicating whether d_i is observed in the given rank list (logging policy) π_q. Joachims *et al.* (2017b) proofed the IPS weighted loss as an unbiased estimate of the loss with true relevance and provided a PBM-based IPS estimator accounting for the position bias. Agarwal *et al.* (2019a) further developed a more general IPS framework with theoretical guarantees by covering more ranking risks/metrics and optimizing neural networks. Building on top of (Wang *et al.*, 2016; Joachims *et al.*, 2017b), a series of counterfactual IPS estimation methods have been developed for correcting selection bias (Ovaisi *et al.*, 2020; Oosterhuis and Rijke, 2020), addressing trust bias (Agarwal *et al.*, 2019c; Vardasbi *et al.*, 2020), mitigating contextual bias (Fang *et al.*, 2019; Chen *et al.*, 2021), etc.

Despite the unbiasedness of propensity-based methods, they may suffer from a high variance issue (Saito, 2020a; Vardasbi *et al.*, 2020; Wang *et al.*, 2021b; Oosterhuis, 2022), potentially due to the irrelevant treatment of non-clicked (displayed) items (Wang *et al.*, 2021b) and the

fact that the inverse values of the propensities could be large (Saito, 2020a). Toward low-variance IPS estimations, Vardasbi *et al.* (2020) applied affine transformation in modeling relevance probability, which not only re-weights the clicks but also penalizes incorrect clicks. For another example, Wang *et al.* (2021b) introduced a ratio-propensity-scoring (RPS) estimator that assigns weights to pairs of clicked and non-clicked items based on the ratio between their propensities, serving as a *biased* estimation yet with low variance.

Doubly Robust Estimation. Doubly-Robust (DR) methods have been widely used in modeling position-biased clicks (Saito, 2020b; Guo *et al.*, 2021; Kiyohara *et al.*, 2022; Zou *et al.*, 2022b). Saito (2020b) introduced a DR method tailored for post-click conversions, followed by a more robust DR estimator (Guo *et al.*, 2021) to reduce variance and a cascade DR estimator (Kiyohara *et al.*, 2022) specifically designed for off-policy evaluation in ranking and search systems. Zou *et al.* (2022b) leveraged DR for relevance estimation by combining low-variance imputation based on large language models and low-bias IPS estimations.

One main obstacle to applying DR in unbiased LTR is the lack of large actual treatments – user examination, which is also one of the main issues in previous IPS approaches since the examine variables (Wang *et al.*, 2016; Joachims *et al.*, 2017b) are not observed in click log history. It would be unclear whether a non-click item was intentionally skipped or not examined by users. To this end, Luo *et al.* (2023) trained a context-aware user simulator with LSTMs to generate pseudo-click labels for unobserved ranking lists. The generated clicked data are incorporated into IPS estimations governed by a doubly robust learning framework, pursuing low-bias and low-variance propensity estimation. Unlike using a neural network-based imputation model (Luo *et al.*, 2023), Oosterhuis (2023) approximated the lacking actual treatment by constructing a covariate with the expectation of treatment per rank and developed a novel DR estimator by jointly optimizing a preference regression model with IPS estimation, resulting in a more robust theoretical guarantee for unbiasedness.

6.3.3 Model-based Methods

Two-Tower Models for learning from biased feedback (Zhao *et al.*, 2019b; Guo *et al.*, 2019; Zhuang *et al.*, 2021; Yan *et al.*, 2022; Zhang *et al.*, 2023) consist of a *relevance tower* and an *observation tower*, where the relevance tower model takes regular input features to predict unbiased relevance, while the observation tower captures the biased-related features inherent in user behaviors (clicks), such as positions and platform (e.g., mobile vs desktop), to estimate non-uniform user observation probability over ranking items. Generally, this kind of method follows the same assumption of PBM models, assuming the relevance prediction and observation probability could be completely factorized, and thus resulting in the popular *two-tower additive model* architecture (Zhao *et al.*, 2019b; Guo *et al.*, 2019; Zhuang *et al.*, 2021). However, recent methods (Yan *et al.*, 2022; Zhang *et al.*, 2023) argue that the independent assumption between relevance and user bias might be too strong in real-world applications. To this end, Yan *et al.* (2022) discussed the limitation of using the additive model to capture user behaviors and enriched user modeling by providing a mixture of EM algorithm and embedding-based interaction. Zhang *et al.* (2023) theoretically showed the confounding effect between relevance and bias models and developed two disentangle methods through gradient reversal and observation dropout.

Neural User Models (Borisov *et al.*, 2016; Zhang *et al.*, 2019; Dai *et al.*, 2020; Luo *et al.*, 2023) seek to predict user behaviors with deep neural networks, which could simulate user data (clicks) and also enrich observations from different ranking lists. Unlike generative click models (Section 6.3.1), which mainly adopt the probabilistic graphical model framework and parameterize user behavior as a sequence of observable and hidden events, the neural click models (Borisov *et al.*, 2016; Borisov *et al.*, 2018) reduce hand-crafted designs and learn to predict user behavior in a data-driven approach, such as training RNNs to represent user clicks with hidden states (Borisov *et al.*, 2016) and adopting an encoder-decoder network to predict user interaction per query session (Borisov *et al.*, 2018). Recently, Zhang *et al.* (2019) developed a context-aware user (click) model to enable a virtual environment for

learning ranking policies via a reinforcement learning framework. Dai *et al.* (2020) trained a deep CTR model to generate counterfactual data for unbiased estimation.

6.4 Evaluation with Biased Relevance Judgments

Accurately evaluating new ranking policy is essential to a wide range of online web search services (Li *et al.*, 2015; Schnabel *et al.*, 2016; Agarwal *et al.*, 2017; Li *et al.*, 2018). Despite various biases existing in user log data (e.g., click, dwell time, etc.), this biased user feedback could provide an inexpensive and fast alternative to unbiased online A/B tests. This section will mainly focus on investigating evaluation with implicit feedback, and further discuss its application in incomplete judgments and fair ranking problems.

6.4.1 Off-policy Evaluation

One mainstream method to perform offline evaluation with implicit feedback is off-policy evaluation (OPE) (Gilotte *et al.*, 2018; Saito and Joachims, 2021; Saito and Joachims, 2022), which aims to realize accurate ranking performance evaluation only using logged data, yet without actual user interactions. A great deal of counterfactual evaluation techniques (Bottou *et al.*, 2013; Li *et al.*, 2015; Swaminathan and Joachims, 2015a) have been developed to implement OPE in a ranking context and to mitigate the distribution shifts between different policies, including model-free estimators, model-based estimators, and the hybrid ones.

Model-Free OPE. The inverse propensity scoring (IPS) estimator (Precup *et al.*, 2000; Strehl *et al.*, 2010; Swaminathan and Joachims, 2015a) forms the standard OPE technique to evaluate rankings (actions) in contextual bandit processes, where the user log data is formulated as contextual bandit feedback from a *logging policy*. The IPS-based methods generally adopt the importance sampling technique to correct the distribution shift between the offline policy and the new online one. The IPS estimator is theoretically guaranteed to be unbiased through two general assumptions of common support and unconfoundedness in causal inference, yet it could be highly vulnerable in large action space

– suffering from high variance. To this end, some advanced IPS variants have been proposed to lower the variance while keeping the unbiased estimation, such as Clipped IPS (Swaminathan and Joachims, 2015a) and Self-Normalized IPS (Swaminathan and Joachims, 2015b), both of which shrink the large IPS weights to balance the bias and variance given in the MSE error. Agarwal *et al.* (2017) also extended IPS to log data obtained from multiple logging policies.

Model-Based OPE. We refer to the model-based methods as two kinds: 1) reward regression models and 2) user behavior models. On the one hand, the direct methods (DM) (Beygelzimer and Langford, 2009) generally optimize a reward regression model to assist policy performance evaluation. While the DM approaches enjoy low variance, they could be highly biased due to less accurate award predictions and model mis-specification (Dudík *et al.*, 2011; Jiang and Li, 2016). On the other hand, different assumptions have been imposed on user behavior models to address the high variance issue of IPS methods, such as the Independent IPS (IIPS) estimator (Li *et al.*, 2018) and the Reward Interaction IPS (RIIPS) (McInerney *et al.*, 2020) estimator, where IIPS implements a new position-level weighted reward function based on the combinatorial action formulation, assuming independent user interactions across positions, and RIIPS adopts a cascade user model – assuming sequential user interactions from top positions to the bottom ones. More recently, an adaptive IPS (AIPS) model (Kiyohara *et al.*, 2023) is proposed to handle diverse user model behaviors by dynamically changing the estimator upon user contexts.

Hybrid OPE. The hybrid methods try to combine the advantages of DM estimators (high bias yet low variance) and IPS estimators (high variance yet low bias) within a doubly robust (DR) estimation framework (Dudík *et al.*, 2011; Jiang and Li, 2016; Kiyohara *et al.*, 2023), which could largely reduce the variance of IPS and remain unbiased. However, the DR estimator, in essence, may still suffer from a high variance issue, especially when it comes to use in a large action space (Saito and Joachims, 2022). To this end, variants of DR (Wang *et al.*, 2017; Su *et al.*, 2019; Su *et al.*, 2020b) have been proposed to achieve better bias-variance control by adjusting the importance weights in the conventional DM-based formulation by, e.g., adaptive weighting (Su

et al., 2019) and carefully hyperparameter turning (Wang *et al.*, 2017; Su *et al.*, 2020b). Kiyohara *et al.* (2022) further incorporated the cascade user model assumption into the DR estimator toward further reducing the variance.

6.4.2 Evaluation with Incomplete Judgments

Convention ranking evaluation follows the Cranfield paradigm (Voorhees, 2002) and employs the complete judgment – the relevance label of every ranked document and the collected document is known. However, it would be more practical to apply unbiased incomplete judgments for evaluation, as annotating all the documents is infeasible and expensive in large-scale search systems (Buckley and Voorhees, 2004; Büttcher *et al.*, 2007). To realize evaluation with incomplete judgments, statistical sampling (Aslam *et al.*, 2006; Yilmaz and Aslam, 2006; Aslam and Yilmaz, 2007) and selection methods (e.g., random selection or top-k pooling (Büttcher *et al.*, 2007)) are two commonly used techniques. Regarding the fairness of ranking, Kırnap *et al.* (2021) recently investigated the impact of incomplete judgments on a series of fair ranking metrics, including both proportion-based (i.e., statistical parity) (Yang and Stoyanovich, 2017; Zehlike *et al.*, 2017) and exposure-based (Singh and Joachims, 2018; Biega *et al.*, 2018) fairness measurements, to alleviate the data-starving and privacy-preserved challenges of labeling attributes in practice. Robust and unbiased estimations were provided to calculate fairness metrics with incomplete judgments based on sampling strategy and the Horvitz-Thompson estimator.

6.4.3 Fair Ranking from Implicit Feedback

Previous fair ranking methods mainly rely on manual relevance judgments for ranking policy training and evaluation (Biega *et al.*, 2018; Singh and Joachims, 2018; Zehlike and Castillo, 2020; Singh and Joachims, 2019), which, however, it is challenging to scale up to large-scale search systems due to the data-starving and privacy challenges. Plus, it has been well recognized that implicit user feedback (e.g., click-through data) could more closely capture user behavior, thus enabling more practical ranking fairness adhering to real-world applications. In

light of this, recent research efforts (Morik *et al.*, 2020; Yadav *et al.*, 2021) have been made in learning fair ranking policy with user click data by solving two main challenges: first, how to incorporate fair constraints into relevance (utility) ranking objectives without using true relevance labels, and second, how to leverage the biased (e.g., position bias) and partial (e.g., user clicks of unexamined items are unknown) implicit user feedback to achieve unbiased relevance and policy estimations.

Morik *et al.* (2020) proposed to simultaneously control bias and fairness within a two-step framework: an unbiased relevance estimator is learned from biased click data first, and the amortized group fairness is then imposed on the learned ranking policy. Following this pioneering work, Yadav *et al.* (2021) further developed a policy-gradient training algorithm to utilize the IPS estimator to learn a better trade-off between fairness and utility from biased user click data. Besides the implicit feedback, several de-biasing techniques have also been explored recently to mitigate the social bias in explicit relevance feedback for fair ranking, such as using data augmentation (Bigdeli *et al.*, 2023) and the label-free distribution-based learning (Chen and Fang, 2023).

6.5 Limitations of Evaluating Fairness

Various evaluation metrics have been proposed to measure the fairness of ranking in search systems (Pitoura *et al.*, 2021; Raj and Ekstrand, 2022). We have demonstrated the details of several commonly used fair ranking metrics in terms of composition, accuracy, and opportunity-based methods (see Table 5.2). In this section, we briefly review the recent works focusing on fair ranking metrics and discuss the potential limitations of fair ranking evaluation.

6.5.1 Fair Ranking Metrics Revisit

Following the definition of fairness in classification, Pitoura *et al.* (2021) summarized an overall taxonomy to specify fair ranking metrics in different levels (*individual fairness* vs. *group fairness*), sides (*user* vs. *item*), and output multiplicity (*single ranking* or *multiple rankings*), which provide a comprehensive understanding of fair ranking and rec-

ommendation toward various aspects of the systems. More recently, Raj and Ekstrand (2022) provided more fine-grained discussions on measuring the fairness of ranking results in the following three aspects:

- *Statistical parity in single rankings* only accesses the exposure equity without measuring the ranking utility – relevance scores, which thus does not require true relevance labels for computing fairness scores. Some representative metrics include prefix fairness (PreF_Δ) (Yang and Stoyanovich, 2017), FAIR (Zehlike *et al.*, 2017), and attention-weighted rank fairness (AWRF_Δ) (Sapiezynski *et al.*, 2019), where Δ denotes different distance functions. PreF_Δ computes statistical parity with position bias by averaging parity over successive prefixes of a ranking list, and FAIR calculates a similarity-based group fairness score based on the top-k positions. AWRF_Δ introduces a position weight model into the fairness metric to better address user behaviors explicitly.

- *Statistical parity in multiple rankings* extends single-ranking evaluation to multiple sequences or distributions of rankings, which is used to evaluate the query-dependent stochastic ranking policy, such as demographic parity (DP) (Singh and Joachims, 2018) and expected exposure disparity (EED) (Diaz *et al.*, 2020). Both DP and EED measure the statistical parity over ranking policies and expect an equal exposure between protected and dominant groups, where DP adopts the ℓ_2 norm to calculate the exposure ratio, and EED captures the inequality in exposure distribution across groups (Raj and Ekstrand, 2022).

- *Equal opportunity in multiple rankings* consider fairness conditioning on ranking utility – the exposure should be proportional to relevance (Singh and Joachims, 2018; Biega *et al.*, 2018; Raj and Ekstrand, 2022). Similar to equality of opportunity in classification (Hardt *et al.*, 2016), Singh and Joachims (2018) developed exposed utility ratio (EUR) and realized utility ratio (RUR) to incorporate rank utility into group fairness exposure. On another hand, the inequity of amortized attention (IAA) metric (Biega *et al.*, 2018) and expected exposure loss (EEL) (Diaz *et al.*, 2020)

have been designed for individual fairness exposure over stochastic ranking policies. Please refer to Section 5.3 for more details about the above equal opportunity metrics.

Raj and Ekstrand (2022) has provided a detailed empirical comparison among the above fair ranking metrics on the TREC Fair Ranking Track 2020 dataset (Biega *et al.*, 2020b). A different perspective is proposed by Gao *et al.* (2022) by combining traditional IR metrics (which are used for assessing relevance) with fairness metrics. More recently, Ratz *et al.* (2024) introduced a new evaluation metric for comparative search result bias based on skewness. Abolghasemi *et al.* (2024) developed an attention-weighted rank fairness evaluation framework over the previous fair ranking metrics (Rekabsaz *et al.*, 2021; Rekabsaz and Schedl, 2020) to access gender bias through the term-based representation of groups in a ranked list.

6.5.2 Open Challenges in Fair Ranking Evaluation

The lack of evaluation benchmarks and tools. While the fairness-aware methods have flourished in recent years, the development of benchmark datasets and evaluation tools fall short in attracting more research efforts, especially for the fair ranking problem, which is often evaluated on synthetic data rather than large-scale real data (Pitoura *et al.*, 2021; Raj and Ekstrand, 2022; Zehlike *et al.*, 2022). To the best of our knowledge, the TREC Fair Ranking Track (Biega *et al.*, 2020b) is the largest public dataset to evaluate the fairness of search systems in ranking documents. Since 2019, the fair tack of TREC has focused on searching relevant academic abstracts from authors belonging to different groups in Semantic Scholar and Wikimedia corpus. However, the type of protected attributes and multi-level attributes are still under-explored. Plus, the evaluation of real-world fair ranking applications (Geyik *et al.*, 2019) is usually inaccessible due to the data privacy issue and the lack of standard access protocols. More evaluation tools, such as FairSearch (Zehlike *et al.*, 2020), IBM's AI Fairness 360, and TensorFlow's Fairness Indicators, are also needed.

The role of relevance feedback. The fair evaluation metrics are generally computed upon manually annotated relevance labels (explicit

feedback), which are expensive to collect and may suffer from cognitive and social bias from assessors. It would be helpful to adopt implicit feedback (e.g., click data) to evaluate ranking fairness, since these user data could largely reduce the acquisition cost and closely describe user behaviors. Balagopalan *et al.* (2023) thoroughly studied the role of relevance in fair ranking and showed it as a good proxy for *worthiness* in fair exposure allocation by providing five validation criteria. The empirical study of click-based relevance in (Balagopalan *et al.*, 2023) casts a new direction of using relevance implicit user feedback in accessing fairness.

Audit the fairness of LLM rankers. Large language models (LLMs) based rankers (Sun *et al.*, 2023; Qin *et al.*, 2023; Ma *et al.*, 2023) have grown rapidly, which generally leverage the generalization and reasoning ability of LLMs to directly "answer" the relevant documents given the query and prompts. However, since the reliability study of large models remains far from mature, there is an urgent need to investigate fair ranking evaluations accounting for the overlarge model size and black-box nature of LLMs. Particularly, how the human-like bias (Ferrara, 2023; Schramowski *et al.*, 2022), misinformation (Nozza *et al.*, 2022), and malicious attack (Wang *et al.*, 2023a) would impact LLM behaviors and further disorder ranking results still needs to be clarified. Scalable validation techniques, calibration data, and mitigation strategies should be designed and developed to audit the fairness of LLM rankers.

7

Research Trends and Future Work

Fairness in search systems is a relatively new but rapidly growing corner of the research literature on information retrieval, machine learning, responsible AI, and related topics. Recent fundamental research breakthroughs at the intersection of the fields of information retrieval, deep learning, large language models, and algorithmic fairness open up fascinating directions for future advancements in the field of fair search.

In this section, we discuss some of these potential directions in detail. We hope that the research presented in this section will inspire our readers to deeply reflect on their assumptions about search and ranking, and eventually advance the field beyond the existing paradigm of search engines, as reflected in current production systems. While information retrieval has a storied past and a well-established present, there is still much that remains to be done to improve fairness in these systems.

7.1 Fairness in Production Ranking Systems

Despite a surge in academic research on developing fair algorithms, their implementation in production search systems used by companies and governments is less prevalent, with limited public disclosure of their fairness strategies. There are noteworthy examples, such as LinkedIn,

where researchers have published their implementation of a system in LinkedIn's recruitment platform that ensures gender-balanced results (Geyik *et al.*, 2019; Quiñonero Candela *et al.*, 2023), representing a user-oriented approach (Li *et al.*, 2023) that aims to deliver fair results directly to end-users. These systems typically require fast response times, necessitating the development of more efficient fair ranking algorithms in future work.

On the other hand, developer-oriented approaches aim to create tools that assist developers and policymakers in the industry to better comprehend and tackle system unfairness. Tools like IBM's AI Fairness 360 (Bellamy *et al.*, 2019), Microsoft's Fairlearn (Bird *et al.*, 2020), and Amazon's SageMaker Clarify (Hardt *et al.*, 2021) exemplify this. They provide ways to identify biases at various points in the machine learning pipeline and offer a range of bias mitigation techniques. Practitioners at Google have also implemented specific fairness metrics in a production ranking system (Beutel *et al.*, 2019b). Such developer-oriented systems often balance multiple objectives, making it challenging to ensure fairness across all aspects of the model and final ranking.

In the realm of recommendation systems, there is notable research, such as Spotify's counterfactual analysis of fairness interventions, balancing user satisfaction with fair artist representation (Mehrotra *et al.*, 2018). Google has also demonstrated improvements in pairwise ranking fairness in their large-scale recommender systems (Beutel *et al.*, 2019a). Moreover, there is research on enhancing fairness in broader machine learning production systems (Bakalar *et al.*, 2021; Madaio *et al.*, 2022; Quiñonero Candela *et al.*, 2023). The work highlights the gap between practical challenges faced by commercial product teams and academic solutions, including issues related to data collection, auditing processes, and human biases. Their insights are equally applicable to search production systems.

7.2 Fairness and Utility

One of the major research challenges is to define fairness and utility in search (Li *et al.*, 2023), largely due to the subjective and context-dependent nature of these concepts. The notion of fairness can vary

across cultures, societies, and individuals. What is considered fair in one context might not be seen as such in another. In addition, search systems serve various stakeholders (providers and consumers), each with their own perceptions of fairness. The multifaceted nature of fairness may lead to a series of research problems. First, current methods primarily address a single type of fairness requirement, even though biases often manifest in multiple forms simultaneously. Consequently, it becomes crucial to explore a unified model that can cater to multiple fairness needs. While achieving a one-size-fits-all solution to all fairness issues is theoretically unattainable; as some fairness definitions are not even compatible except in highly constrained special cases (Kleinberg *et al.*, 2016) — it is still valuable to explore some simpler scenarios such as handling two or three different cases. Secondly, an important challenge arises when certain fairness requirements conflict and cannot be achieved simultaneously. The key issue here is how to establish a reasonable and effective trade-off. Recent advancement in fairness research in classification (Hsu *et al.*, 2022) has demonstrated the potential to meet all fairness criteria with minimal violations. Extending this approach to fair search presents an intriguing area for research.

Similarly, defining the utility of a search system can be challenging too (Patro *et al.*, 2022). The fair ranking literature often employs exposure as a surrogate for provider utility, exemplified by concepts like fairness of exposure or equity of attention (Biega *et al.*, 2018; Singh and Joachims, 2018), as discussed in Section 5. These approaches typically assume that exposure correlates directly with a provider's ranking position, where each position is assigned a fixed value, irrespective of context. This ranking-based perspective on exposure might overlook important context-specific factors. For instance, higher exposure does not always equate to increased user attention, and even when it does, this heightened attention does not necessarily translate into tangible provider utility, such as sales or long-term satisfaction. An additional critical contextual factor that extends beyond mere ranking position is time, particularly in rapidly evolving domains like news, where items retain relevance for only a brief period (Campos *et al.*, 2014). In these dynamic environments, both users and providers derive the most benefit from immediate exposure. For instance, the timeliness of information is a

key component of relevance in breaking news (Chakraborty *et al.*, 2017). Another example is that restaurants are likely to receive more orders if they are promoted in ranking during peak hours to customers in close proximity (Banerjee *et al.*, 2020). How to consider all the contextual factors and define a more suitable and realistic utility function remains an important area for research.

Balancing fairness and utility poses a complex challenge, as improving one may impact the other. Striking the right balance requires ongoing research, development, and ethical considerations. In many industries, businesses prioritize metrics such as purchase rates over treating their users fairly. As a result, the motivation to promote fairness is often overshadowed by the pursuit of profits, particularly when there is a trade-off between fairness and profit metrics. However, legal requirements in many countries, such as the GDPR in the EU[1], CCPA in the US[2], and IISARR in China[3], enforce fair treatment of users in practical systems. It is crucial for the research community to examine the relationship between fairness and utility to encourage industry practitioners to prioritize fairness. In some cases, fairness and utility can reinforce each other. For example, a fair set of search results can provide a more comprehensive view of a topic, enhancing utility. Some work on classification tasks has also found that improving fairness may improve overall accuracy (Lahoti *et al.*, 2020). In case certain measures of fairness and utility are inherently at odds with each other, we can create approaches that ensure fairness with minimal impact on utility, or ensure utility with minimal impact on fairness. These impacts can be explained to users in a suitable manner so that they can comprehend and accept them, and distributed across users or time to avoid continuously harming particular groups of users.

7.3 Data and Benchmarks

While Section 5 introduced several benchmark datasets, the availability of data tailored for fair search research is notably limited. This scarcity

[1] https://gdpr.eu/
[2] https://oag.ca.gov/privacy/ccpa
[3] http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm

is partly due to the unique requirements of fairness research, which often necessitates datasets containing user personal information, such as gender, race, age, location, and income. Moreover, the multifaceted nature of fairness concerns requires a diverse range of datasets, each catering to specific aspects of fairness.

Compiling such comprehensive datasets poses significant challenges, particularly from a legal perspective. For instance, the data minimization principle of the GDPR restricts the collection of sensitive information like gender or race, which could inadvertently impede the implementation of fairness interventions. Reliance on inferred attributes introduces considerable uncertainty, potentially diminishing the effectiveness of these interventions. Biega *et al.* (2020a) suggested that while data minimization might not drastically reduce performance, it could disproportionately affect different user groups.

Legal hurdles extend beyond privacy concerns to include data retention policies and intellectual property rights of platforms. To navigate these challenges, robust anonymization of users and innovative methods like federated learning (Kairouz *et al.*, 2021) or differential privacy (Dwork, 2006) can be explored for balancing privacy with fairness. Concurrently, regulatory efforts are encouraging greater transparency in algorithmic systems. The Federal Trade Commission's Algorithmic Accountability Act in the U.S., for example, mandates external consultations for impact assessments, involving independent auditors and technology experts, to ensure impartial evaluations of platform-operated systems (Gursoy *et al.*, 2022).

In scenarios where direct access to data or full knowledge of ranking algorithms is unavailable, researchers can still leverage simulations. Tools like Virtual-Taobao (Shi *et al.*, 2019) and AESim (Gao *et al.*, 2021c) have been developed for this purpose, simulating the interaction dynamics between stakeholders and the system under certain assumptions. It is crucial for these simulation frameworks to remain flexible, allowing researchers to modify or choose the foundational assumptions to more accurately reflect real-world search environments.

7.4 Causal Fairness

The current focus in quantifying fairness in search systems is predominantly on statistical-based measures, which assess correlations between predictive outcomes and sensitive attributes. These methods are preferred for their relative ease of calculation and implementation. However, they have limitations, as they rely on correlation rather than causation. This means they can only determine fairness based on the specific metric being used. Furthermore, addressing any identified fairness disparities requires not only an understanding of how these statistics are generated but also insights into assigning responsibility and devising remedies.

Looking at fairness from legal and philosophical perspectives uncovers another drawback of statistical fairness notions. In discrimination cases, for instance, there might be a need to demonstrate a causal link between the outcome (such as ranking of candidates in hiring) and a sensitive attribute (such as gender or race). This necessitates an exploration of the causal relationships rather than just associative ones.

Causal fairness metrics, which measure the causal effects of sensitive features on outcomes, offer a more nuanced analysis. They examine the dependency between protected attributes and final decisions, allowing for an investigation into the actual causes of unfairness, which statistical measures cannot provide. In search systems, this means addressing not just the symptoms, but the root causes of unfairness. For example, causal fairness seeks to identify and correct structural and algorithmic biases, focusing on protected attributes like race, gender, or age.

Contrary to statistical-based approaches that rely solely on data, causal-based fairness incorporates additional structural knowledge of how variables interact in a causal model, like a causal graph (Makhlouf *et al.*, 2020). These notions typically involve interventions and counterfactuals (Li *et al.*, 2023; Le Quy *et al.*, 2022), and promising approaches have been proposed for classification (Kusner *et al.*, 2017) and recommendation systems (Li *et al.*, 2021a). The application of causality in fair search is still nascent (Yang *et al.*, 2020), but we anticipate that incorporating causal considerations will introduce new challenges and opportunities for advancing fairness in search systems.

7.5 Large Language Models and Search

The advent of Large Language Models (LLMs) such as GPT models (OpenAI, 2023) and LLaMA (Touvron *et al.*, 2023) has revolutionized the field of natural language processing, as these transformer-based models exhibit remarkable language understanding, generation, and generalization capabilities as their sizes are scaled up. Despite their capabilities, LLMs come with their own set of limitations, such as hallucination, lack of commonsense reasoning, and sociotechnical concerns related to fairness and bias. This is an emerging area of research with only preliminary findings to date.

Recent research has sought to leverage LLMs to improve various components in a search system, including query rewriters, retrievers, rerankers, and readers, demonstrating promising results (Zhu *et al.*, 2023). As LLMs gain prominence, assessing their fairness is becoming as imperative as evaluating their effectiveness and efficiency, especially given their widespread impact and accessibility. Previous studies in NLP (Hutchinson *et al.*, 2020; Perez *et al.*, 2022; Abid *et al.*, 2021) have documented instances of language models exhibiting bias against marginalized groups.

While the fairness of traditional search engines has been considerably investigated, there remains a significant research gap in understanding how LLMs, when employed as components or entire systems of IR, impact fairness in search. Recently, Dai *et al.* (2024a) presented a survey on bias and unfairness in IR systems integrated with LLMs. Their work frames bias and unfairness as distribution mismatch problems and examines specific issues arising at three stages of LLM integration into IR systems: data collection, model development, and result evaluation. Notably, much of the cited research focuses on recommendation systems rather than search engines. Furthermore, Wang *et al.* (2024b) evaluated fairness in ranking tasks with LLMs, and Wu *et al.* (2024b), Hu *et al.* (2024), and Kim and Diaz (2024) investigated fairness in Retrieval-Augmented Generation (RAG) systems. As LLMs are increasingly employed to automatically evaluate search systems (Thomas *et al.*, 2024; Rahmani *et al.*, 2024), to generate relevance labels (Khramtsova *et al.*, 2024; Zhuang *et al.*, 2024; Zhang *et al.*, 2024a), and to annotate

group membership for group fairness assessments (Chen *et al.*, 2024), the extent to which these automated evaluations introduce new fairness issues remains unclear.

Additionally, LLMs or foundation models at large have been instrumental in the creation of high-quality AI Generated Content (AIGC). These models, with their expansive capabilities, facilitate the rapid development of domain-specific models commonly used for generating diverse content types, including text, images, audio, and video. For instance, Stable Diffusion (Yang *et al.*, 2023b) and DALL-E 3 (Betker *et al.*, 2023) can produce high-quality images from brief text descriptions. The growing application of AIGC across content production pipelines in society and the Web introduces potential risks. Studies have identified concerns with AI-generated content, including issues related to discrimination and representational harms (Jiang *et al.*, 2023; Deshpande *et al.*, 2023).

The training data for Generative AI models, being sourced from the real world, may inadvertently perpetuate harmful stereotypes, overlook or marginalize certain groups, and include toxic data sources. This can lead to content that promotes discrimination or hate (Weidinger *et al.*, 2021; Birhane *et al.*, 2021). While addressing biases and stereotypes at the source data level is a step forward, it is essential to assess unfairness throughout the entire model training and development life cycle, extending beyond just the data source. Furthermore, defining what constitutes a truly unbiased dataset poses a significant challenge. The depth and characteristics of these issues within Generative AI models have not yet been comprehensively investigated (Chen *et al.*, 2023a).

A key question that emerges as the Internet is increasingly populated with AIGC is its impact on the ranking results of retrieval systems. In exploring this, Dai *et al.* (2024b) revealed that neural IR models exhibit a bias towards text generated by LLMs, a phenomenon they term as *source bias*. Building on this, Xu *et al.* (2023) extended the study of source bias in AIGC to include text-image retrieval models. Their findings indicate that these models often rank AI-generated images higher than real images, despite the AI-generated images not necessarily exhibiting more visually relevant features to the query than real images. This

form of invisible relevance bias is widespread, affecting various retrieval models with different training data and architectures. Moreover, their research suggests that including AI-generated images in the training data of retrieval models further intensifies this invisible relevance bias.

Anthis *et al.* (2024) examined the applications of existing fairness frameworks such as group fairness and fair representations to LLMs and showed that these frameworks either do not logically extend to LLMs or present a notion of fairness intractable for LLMs, largely due to the multitudes of populations affected, sensitive attributes, and use cases. Despite the challenges, they demonstrated feasibility for achieving fairness in particular use cases with guidelines. In light of these existing studies, an important research direction lies in developing strategies to mitigate the unfairness observed in LLMs, particularly within the context of search ecosystems.

7.6 Concluding Remarks

As pointed out in Section 1.5, this monograph aims to serve as an entry point to the field of fairness in search systems, designed to be accessible to a broad audience, including those with backgrounds in information retrieval and AI ethics. The rise of LLMs is profoundly transforming search technologies, presenting new and complex challenges in the pursuit of mitigating bias and enhancing fairness. In addition, ongoing research in responsible AI is continually enriching our understanding of fairness in modern information systems. Although we have not covered this rapidly developing field in its entirety, we have aimed to provide a holistic overview and framework by compiling and synthesizing existing research to facilitate a deeper understanding and help address these emerging issues. As the field evolves, new areas and challenges will undoubtedly emerge, and we look forward to the ongoing development of this field, which is poised to advance our grasp of fairness in search systems.

Acknowledgements

This work was partially supported by DOCOMO Innovations, the Ciocca Center research award, and the Responsible AI program at Santa Clara University. We are grateful to the journal editors and anonymous reviewers whose feedback and suggestions greatly enhanced this manuscript. Thanks to Xuyang Wu, Zhiyuan Peng, and Leo Wei for their diligent proofreading efforts.

References

Abdollahpouri, H., G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. (2019). "Beyond personalization: Research directions in multistakeholder recommendation". *arXiv preprint arXiv:1905.01986*.

Abdollahpouri, H. and R. Burke. (2019). "Multi-stakeholder recommendation and its connection to multi-sided fairness". *arXiv preprint arXiv:1907.13158*.

Abdollahpouri, H., R. Burke, and B. Mobasher. (2017). "Controlling popularity bias in learning-to-rank recommendation". In: *RecSys*.

Abid, A., M. Farooqi, and J. Zou. (2021). "Large language models associate Muslims with violence". *Nature Machine Intelligence*. 3(June): 461–463. DOI: 10.1038/s42256-021-00359-2.

Abolghasemi, A., L. Azzopardi, A. Askari, M. de Rijke, and S. Verberne. (2024). "Measuring Bias in a Ranked List Using Term-Based Representations". In: *European Conference on Information Retrieval*. Springer. 3–19.

Agarwal, A., K. Takatsu, I. Zaitsev, and T. Joachims. (2019a). "A General Framework for Counterfactual Learning-to-Rank". In: *SIGIR*.

Agarwal, A., I. Zaitsev, X. Wang, C. Li, M. Najork, and T. Joachims. (2019b). "Estimating Position Bias Without Intrusive Interventions". In: *WSDM*.

Agarwal, A., S. Basu, T. Schnabel, and T. Joachims. (2017). "Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery. 687–696.

Agarwal, A., X. Wang, C. Li, M. Bendersky, and M. Najork. (2019c). "Addressing Trust Bias for Unbiased Learning-to-Rank". In: *The World Wide Web Conference*. 4–14.

Agarwal, A., I. Zaitsev, X. Wang, C. Li, M. Najork, and T. Joachims. (2019d). "Estimating Position Bias without Intrusive Interventions". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 474–482.

Ahmad, W. U., K.-W. Chang, and H. Wang. (2018). "Multi-Task Learning for Document Ranking and Query Suggestion". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJ1nzBeA-.

Ai, Q., J. Mao, Y. Liu, and W. B. Croft. (2018). "Unbiased Learning to Rank: Theory and Practice". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. Torino, Italy: Association for Computing Machinery. 2305–2306.

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. (2022). "Flamingo: a Visual Language Model for Few-Shot Learning". In: *Advances in Neural Information Processing Systems*.

Ali, M., P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke. (2019). "Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes". *CSCW*.

Angwin, J. and J. Larson. (2016a). "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say". URL: https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

Angwin, J. and J. Larson. (2016b). "How We Analyzed the COMPAS Recidivism Algorithm". URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Anthis, J., K. Lum, M. Ekstrand, A. Feller, A. D'Amour, and C. Tan. (2024). "The Impossibility of Fair LLMs". In: *Proceedings of the 1st Human-Centered Evaluation and Auditing of Language Models (HEAL) workshop at CHI 2024*.

Aslam, J. A., V. Pavlu, and E. Yilmaz. (2006). "A Statistical Method for System Evaluation Using Incomplete Judgments". In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06*. Seattle, Washington, USA: Association for Computing Machinery. 541–548.

Aslam, J. A. and E. Yilmaz. (2007). "Inferring Document Relevance from Incomplete Information". In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. CIKM '07*. Lisbon, Portugal: Association for Computing Machinery. 633–642.

Asudeh, A., H. Jagadish, J. Stoyanovichz, and G. Das. (2019). "Designing Fair Ranking Schemes". *ICDM*.

Azzopardi, L. (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*. Association for Computing Machinery. 27–37.

Azzopardi, L. and V. Vinay. (2008). "Retrievability: An evaluation measure for higher order information access tasks". In: *Proceedings of the 17th ACM conference on Information and knowledge management*. 561–570.

Badilla, P., F. Bravo-Marquez, and J. Pérez. (2020). "WEFE: The Word Embeddings Fairness Evaluation Framework". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by C. Bessiere. International Joint Conferences on Artificial Intelligence Organization. 430–436. DOI: 10.24963/ijcai.2020/60.

Baeza-Yates, R. (2018). "Bias on the web". *Communications of the ACM*. 61(6): 54–61.

Bagdasaryan, E., O. Poursaeed, and V. Shmatikov. (2019). "Differential privacy has disparate impact on model accuracy". *Advances in neural information processing systems*. 32.

Bakalar, C., R. Barreto, S. Bergman, M. Bogen, B. Chern, S. Corbett-Davies, M. Hall, I. Kloumann, M. Lam, J. Q. Candela, *et al.* (2021). "Fairness on the ground: Applying algorithmic fairness approaches to production systems". *arXiv preprint arXiv:2103.06172*.

Balagopalan, A., A. Z. Jacobs, and A. J. Biega. (2023). "The Role of Relevance in Fair Ranking". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. New York, NY, USA: Association for Computing Machinery. 2650–2660.

Balaneshin-kordan, S. and A. Kotov. (2017). "Embedding-Based Query Expansion for Weighted Sequential Dependence Retrieval Model". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17*. Shinjuku, Tokyo, Japan: Association for Computing Machinery. 1213–1216.

Banerjee, A., G. K. Patro, L. W. Dietz, and A. Chakraborty. (2020). "Analyzing 'Near Me'Services: Potential for Exposure Bias in Location-based Retrieval". In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 3642–3651.

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

Barocas, S. and A. D. Selbst. (2016). "Big data's disparate impact". *California law review*: 671–732.

Bashir, S. and A. Rauber. (2010). "Improving retrievability of patents in prior-art search". In: *European Conference on Information Retrieval*. Springer. 457–470.

Bellamy, R. K., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, *et al.* (2019). "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". *IBM Journal of Research and Development*. 63(4/5): 4–1.

Bellogin, A., L. Boratto, S. Kleanthous, E. Lex, F. M. Malloci, and M. Marras. (2024). "International Workshop on Algorithmic Bias in Search and Recommendation (BIAS)". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3033–3035.

Bengio, Y., A. Courville, and P. Vincent. (2013). "Representation Learning: A Review and New Perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 35(8): 1798–1828. DOI: 10.1109/TPAMI.2013.50.

Bernard, N. and K. Balog. (2023). "A Systematic Review of Fairness, Accountability, Transparency and Ethics in Information Retrieval". *ACM Computing Surveys*.

Betker, J., G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, *et al.* (2023). "Improving image generation with better captions". *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*.

Beutel, A., J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. (2019a). "Fairness in Recommendation Ranking through Pairwise Comparisons". In: *KDD*.

Beutel, A., J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. (2019b). "Putting fairness principles into practice: Challenges, metrics, and improvements". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

Beutel, A., J. Chen, Z. Zhao, and E. H. Chi. (2017). "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations". *ArXiv*. abs/1707.00075. URL: https://api.semanticscholar.org/CorpusID:24990444.

Beygelzimer, A. and J. Langford. (2009). "The Offset Tree for Learning with Partial Labels". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09*. Paris, France: Association for Computing Machinery. 129–138.

Biega, A., F. Diaz, M. Ekstrand, and S. Kohlmeier. (2019). "TREC 2019 Fair Ranking Track".

Biega, A. J., K. P. Gummadi, and G. 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.

Biega, A. J., P. Potash, H. Daume, F. Diaz, and M. Finck. (2020a). "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.

Biega, A. J., F. Diaz, M. D. Ekstrand, and S. Kohlmeier. (2020b). "Overview of the TREC 2019 Fair Ranking Track". *CoRR*. abs/2003.11650. URL: https://arxiv.org/abs/2003.11650.

Bigdeli, A., N. Arabzadeh, S. Seyedsalehi, B. Mitra, M. Zihayat, and E. Bagheri. (2023). "De-Biasing Relevance Judgements for Fair Ranking". In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*. Dublin, Ireland: Springer-Verlag. 350–358. DOI: 10.1007/978-3-031-28238-6_24.

Bigdeli, A., N. Arabzadeh, S. SeyedSalehi, M. Zihayat, and E. Bagheri. (2022). "Gender Fairness in Information Retrieval Systems". In: *SIGIR*.

Bigdeli, A., N. Arabzadeh, S. Seyedsalehi, M. Zihayat, and E. Bagheri. (2021a). "On the orthogonality of bias and utility in ad hoc retrieval". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1748–1752.

Bigdeli, A., N. Arabzadeh, M. Zihayat, and E. Bagheri. (2021b). "Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets". In: *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*. Springer-Verlag. 216–224.

Bird, S., M. Dudik, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. (2020). "Fairlearn: A toolkit for assessing and improving fairness in AI". *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Birhane, A., V. U. Prabhu, and E. Kahembwe. (2021). "Multimodal datasets: misogyny, pornography, and malignant stereotypes". *arXiv preprint arXiv:2110.01963*.

Bithel, S. and S. Bedathur. (2023). "Evaluating Cross-Modal Generative Models Using Retrieval Task". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. New York, NY, USA: Association for Computing Machinery. 1960–1965.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. (2016a). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Bolukbasi, T., K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. (2016b). "Quantifying and Reducing Stereotypes in Word Embeddings". *CoRR*. abs/1606.06121. arXiv: 1606.06121. URL: http://arxiv.org/abs/1606.06121.

Bonart, M., A. Samokhina, G. Heisenberg, and P. Schaer. (2020). "An investigation of biases in web search engine query suggestions". *Online Information Review*. 44(2): 365–381.

Boratto, L., S. Faralli, M. Marras, and G. Stilo. (2023). "Fourth International Workshop on Algorithmic Bias in Search and Recommendation (Bias 2023)". In: *European Conference on Information Retrieval*. Springer. 373–376.

Borisov, A., I. Markov, M. de Rijke, and P. Serdyukov. (2016). "A Neural Click Model for Web Search". In: *Proceedings of the 25th International Conference on World Wide Web*. 531–541.

Borisov, A., M. Wardenaar, I. Markov, and M. de Rijke. (2018). "A Click Sequence Model for Web Search". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: Association for Computing Machinery. 45–54.

Bottou, L., J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. (2013). "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising". *Journal of Machine Learning Research*. 14(101): 3207–3260. URL: http://jmlr.org/papers/v14/bottou13a.html.

Boyce, B. (1982). "Beyond topicality: A two stage view of relevance and the retrieval process". *Information Processing & Management*. 18(3): 105–109.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* (2020). "Language models are few-shot learners". *Advances in neural information processing systems*. 33: 1877–1901.

Brunet, M.-E., C. Alkalay-Houlihan, A. Anderson, and R. Zemel. (2019). "Understanding the Origins of Bias in Word Embeddings". In: *Proceedings of the 36th International Conference on Machine Learning*. 803–811.

Buckley, C. and E. M. Voorhees. (2004). "Retrieval Evaluation with Incomplete Information". In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04*. Sheffield, United Kingdom: Association for Computing Machinery. 25–32.

Buolamwini, J. and T. Gebru. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 77–91.

Burke, R., N. Sonboli, and A. Ordonez-Gauger. (2018). "Balanced Neighborhoods for Multi-sided Fairness in Recommendation". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*.

Büttcher, S., C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. (2007). "Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07*. Amsterdam, The Netherlands: Association for Computing Machinery. 63–70.

Cachel, K. and E. Rundensteiner. (2024). "Wise Fusion: Group Fairness Enhanced Rank Fusion". In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 163–174.

Cai, M., H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D. Park, and Y. J. Lee. (2024). "Making Large Multimodal Models Understand Arbitrary Visual Prompts". In: *IEEE Conference on Computer Vision and Pattern Recognition*.

Calders, T., F. Kamiran, and M. Pechenizkiy. (2009). "Building classifiers with independency constraints". In: *Data mining workshops, ICDMW*. 13–18.

Caliskan, A., J. J. Bryson, and A. Narayanan. (2017). "Semantics derived automatically from language corpora contain human-like biases". *Science*. 356(6334): 183–186. DOI: 10.1126/science.aal4230. eprint: https://www.science.org/doi/pdf/10.1126/science.aal4230.

Calmon, F., D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. (2017). "Optimized pre-processing for discrimination prevention". *Advances in neural information processing systems*. 30.

Campos, P. G., F. Diez, and I. Cantador. (2014). "Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols". *User Modeling and User-Adapted Interaction*. 24: 67–119.

Carbonell, J. and J. Goldstein. (1998). "The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries". In: *SIGIR*. Melbourne, Australia. 335–336. DOI: 10.1145/290941.291025.

Castelnovo, A., R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini. (2022). "A clarification of the nuances in the fairness metrics landscape". *Scientific Reports*. 12(1): 4209.

Caton, S. and C. Haas. (2020). "Fairness in machine learning: A survey". *ACM Computing Surveys*.

Celis, L. E., D. Straszak, and N. K. Vishnoi. (2018). "Ranking with Fairness Constraints". *ICALP*.

Celma, Ò. and P. Cano. (2008). "From hits to niches?: or how popular artists can bias music recommendation and discovery". In: *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*. ACM. 5.

Chakraborty, A., S. Ghosh, N. Ganguly, and K. 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.

Chaney, A. J., B. M. Stewart, and B. E. Engelhardt. (2018). "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility". In: *Proceedings of the 12th ACM conference on recommender systems*. 224–232.

Chapelle, O. and Y. Zhang. (2009). "A Dynamic Bayesian Network Click Model for Web Search Ranking". In: *Proceedings of the 18th International Conference on World Wide Web*. 1–10.

Chen, C., J. Fu, and L. Lyu. (2023a). "A pathway towards responsible ai generated content". *arXiv preprint arXiv:2303.01325*.

Chen, F. and H. Fang. (2023). "Learn to Be Fair without Labels: A Distribution-Based Learning Framework for Fair Ranking". In: *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '23*. Taipei, Taiwan: Association for Computing Machinery. 23–32.

Chen, F., D. Yang, and H. Fang. (2024). "Toward Automatic Group Membership Annotation for Group Fairness Evaluation". In: *International Conference on Applications of Natural Language to Information Systems*. Springer. 285–300.

Chen, J., H. Dong, X. Wang, F. Feng, M. Wang, and X. He. (2023b). "Bias and debias in recommender system: A survey and future directions". *ACM Transactions on Information Systems*. 41(3): 1–39.

Chen, M., C. Liu, J. Sun, and S. C. Hoi. (2021). "Adapting Interactional Observation Embedding for Counterfactual Learning to Rank". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 285–294.

Cherumanal, S. P., F. Scholer, J. R. Trippas, and D. Spina. (2024). "Towards Investigating Biases in Spoken Conversational Search". *arXiv preprint arXiv:2409.00890*.

Cho, J., A. Zala, and M. Bansal. (2023). "DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models". In: *ICCV*.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

Cho, S., K. W. Crenshaw, and L. McCall. (2013). "Toward a field of intersectionality studies: Theory, applications, and praxis". *Signs: Journal of women in culture and society*. 38(4): 785–810.

Chouldechova, A. (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". *Big data*. 5(2): 153–163.

Chuklin, A., I. Markov, and M. de Rijke. (2015). "Click Models for Web Search". In: *Click Models for Web Search*. URL: https://api.semanticscholar.org/CorpusID:38886570.

Clarke, C. L., M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. (2008). "Novelty and Diversity in Information Retrieval Evaluation". In: *SIGIR*. Singapore, Singapore. 659–666. DOI: 10.1145/1390334.1390446.

Commission, E. (2020). "Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150". URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020XC1208(01)&from=EN.

Crane, D. A. (2011). "Search neutrality as an antitrust principle". *Geo. Mason L. Rev.* 19: 1199.

Craswell, N., O. Zoeter, M. Taylor, and B. Ramsey. (2008a). "An Experimental Comparison of Click Position-Bias Models". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining. WSDM '08*. Palo Alto, California, USA: Association for Computing Machinery. 87–94.

Craswell, N., O. Zoeter, M. Taylor, and B. Ramsey. (2008b). "An experimental comparison of click position-bias models". In: *WSDM*.

Crawford, K. (2017). "The trouble with bias (Invited Talk)". In: *NIPS*.

Dai, S., C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu. (2024a). "Bias and unfairness in information retrieval systems: New challenges in the llm era". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6437–6447.

Dai, S., Y. Zhou, L. Pang, W. Liu, X. Hu, Y. Liu, X. Zhang, G. Wang, and J. Xu. (2024b). "Neural retrievers are biased towards llm-generated content". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 526–537.

Dai, X., J. Hou, Q. Liu, Y. Xi, R. Tang, W. Zhang, X. He, J. Wang, and Y. Yu. (2020). "U-Rank: Utility-Oriented Learning to Rank with Implicit Feedback". In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 2373–2380.

Dai, Z., V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. Hall, and M.-W. Chang. (2023). "Promptagator: Few-shot Dense Retrieval From 8 Examples". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=gmL46YMpu2J.

Dambanemuya, H. K. and N. Diakopoulos. (2021). "Auditing the Information Quality of News-Related Queries on the Alexa Voice Assistant". *Proceedings of the ACM on Human-Computer Interaction*. 5(CSCW1): 1–21.

Dash, A., A. Chakraborty, S. Ghosh, A. Mukherjee, and K. P. Gummadi. (2022). "Alexa, in you, I trust! Fairness and Interpretability Issues in E-commerce Search through Smart Speakers". In: *Proceedings of the ACM Web Conference 2022*. 3695–3705.

Deshpande, A., V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan. (2023). "Toxicity in chatgpt: Analyzing persona-assigned language models". *arXiv preprint arXiv:2304.05335*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

Diaz, F., B. Mitra, and N. Craswell. (2016). "Query Expansion with Locally-Trained Word Embeddings". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by K. Erk and N. A. Smith. Berlin, Germany: Association for Computational Linguistics. 367–377. DOI: 10.18653/v1/P16-1035.

Diaz, F., B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. (2020). "Evaluating stochastic rankings with expected exposure". In: *CIKM*.

Drosou, M., H. V. Jagadish, E. Pitoura, and J. Stoyanovich. (2017). "Diversity in big data: A review". *Big data*. 5(2): 73–84.

Dudík, M., J. Langford, and L. Li. (2011). "Doubly Robust Policy Evaluation and Learning". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning. ICML'11*. Bellevue, Washington, USA: Omnipress. 1097–1104.

Dupret, G. E. and B. Piwowarski. (2008). "A User Browsing Model to Predict Search Engine Click Data from Past Observations." In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '08*. Singapore, Singapore: Association for Computing Machinery. 331–338.

Dwork, C. (2006). "Differential privacy". In: *International colloquium on automata, languages, and programming*. Springer. 1–12.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. (2012). "Fairness through awareness". In: *ITCS*. 214–226.

Dwork, C., M. P. Kim, O. Reingold, G. N. Rothblum, and G. Yona. (2019). "Learning from outcomes: Evidence-based rankings". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 106–125.

Eickhoff, C. (2018). "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18*. Marina Del Rey, CA, USA: Association for Computing Machinery. 162–170.

Ekstrand, M. D., A. Das, R. Burke, F. Diaz, *et al.* (2022). "Fairness in information access systems". *Foundations and Trends® in Information Retrieval*. 16(1-2): 1–177.

Ekstrand, M. D., M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. 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.

Fang, H. and P. Xie. (2020). "CERT: Contrastive Self-supervised Learning for Language Understanding". *CoRR*. abs/2005.12766.

Fang, Y., H. Liu, Z. Tao, and M. Yurochkin. (2022). "Fairness of Machine Learning in Search Engines". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 5132–5135.

Fang, Z., A. Agarwal, and T. Joachims. (2019). "Intervention Harvesting for Context-Dependent Examination-Bias Estimation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19*. Paris, France: Association for Computing Machinery. 825–834.

Feldman, M., S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. (2015). "Certifying and removing disparate impact". In: *KDD*.

Ferrara, E. (2023). "Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models". *CoRR*. abs/2304.03738. DOI: 10.48550/arXiv.2304.03738. arXiv: 2304.03738.

Fleder, D. and K. Hosanagar. (2009). "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity". *Management science*. 55(5): 697–712.

Foulds, J. R., R. Islam, K. N. Keya, and S. Pan. (2020). "An intersectional definition of fairness". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 1918–1921.

Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian. (2021). "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making". *Communications of the ACM*. 64(4): 136–143.

Friedman, B. and H. Nissenbaum. (1996). "Bias in computer systems". *ACM Transactions on information systems (TOIS)*. 14(3): 330–347.

Gallegos, I. O., R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. (2023). "Bias and Fairness in Large Language Models: A Survey". arXiv: 2309.00770 [cs.CL].

Ganguly, D., D. Roy, M. Mitra, and G. J. Jones. (2015). "Word Embedding Based Generalized Language Model for Information Retrieval". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. Santiago, Chile: Association for Computing Machinery. 795–798.

Gao, R., Y. Ge, and C. Shah. (2022). "FAIR: Fairness-aware information retrieval evaluation". *Journal of the Association for Information Science and Technology*. 73(10): 1461–1473.

Gao, R. and C. Shah. (2019). "How fair can we go: Detecting the boundaries of fairness optimization in information retrieval". In: *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*. 229–236.

Gao, R. and C. Shah. (2020). "Toward creating a fairer ranking in search engine results". *Information Processing & Management*. 57(1): 102138.

Gao, R. and C. Shah. (2021). "Addressing bias and fairness in search systems". In: *SIGIR*.

Gao, T., A. Fisch, and D. Chen. (2021a). "Making Pre-trained Language Models Better Few-shot Learners". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3816–3830.

Gao, T., A. Fisch, and D. Chen. (2021b). "Making Pre-trained Language Models Better Few-shot Learners". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics. 3816–3830. DOI: 10.18653/v1/2021.acl-long.295.

Gao, Y., G. Huzhang, W. Shen, Y. Liu, W.-J. Zhou, Q. Da, and Y. Yu. (2021c). "Imitate TheWorld: A Search Engine Simulation Platform". *arXiv preprint arXiv:2107.07693*.

Gerritse, E. J., F. Hasibi, and A. P. de Vries. (2020). "Bias in conversational search: The double-edged sword of the personalized knowledge graph". In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 133–136.

Geyik, S. C., S. Ambler, and K. Kenthapadi. (2019). "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". *KDD*.

Ghazimatin, A., M. Kleindessner, C. Russell, Z. Abedjan, and J. Golebiowski. (2022). "Measuring fairness of rankings under noisy sensitive information". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2263–2279.

Gilotte, A., C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. (2018). "Offline A/B Testing for Recommender Systems". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18*. Marina Del Rey, CA, USA: Association for Computing Machinery. 198–206.

Goldfarb-Tarrant, S., R. Marchant, R. Muñoz Sánchez, M. Pandya, and A. Lopez. (2021). "Intrinsic Bias Metrics Do Not Correlate with Application Bias". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics. 1926–1940. DOI: 10.18653/v1/2021.acl-long.150.

Gomroki, G., H. Behzadi, R. Fattahi, and J. S. Fadardi. (2023). "Identifying effective cognitive biases in information retrieval". *Journal of Information Science*. 49(2): 348–358.

Gonen, H. and Y. Goldberg. (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of the 2019 Workshop on Widening NLP*. Ed. by A. Axelrod, D. Yang, R. Cunha, S. Shaikh, and Z. Waseem. Florence, Italy: Association for Computational Linguistics. 60–63. URL: https://aclanthology.org/W19-3621.

Gorti, S. K., N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, and G. Yu. (2022). "X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5006–5015.

Gu, Y., W. Bao, D. Ou, X. Li, B. Cui, B. Ma, H. Huang, Q. Liu, and X. Zeng. (2021). "Self-Supervised Learning on Users' Spontaneous Behaviors for Multi-Scenario Ranking in E-Commerce". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management. CIKM '21*. 3828–3837.

Guo, F., C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos. (2009a). "Click Chain Model in Web Search". In: *Proceedings of the 18th International Conference on World Wide Web. WWW '09*. Madrid, Spain: Association for Computing Machinery. 11–20.

Guo, F., C. Liu, and Y. M. Wang. (2009b). "Efficient Multiple-Click Models in Web Search". In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining. WSDM '09*. Barcelona, Spain: Association for Computing Machinery. 124–131.

Guo, H., J. Yu, Q. Liu, R. Tang, and Y. 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. RecSys '19*. Copenhagen, Denmark: Association for Computing Machinery. 452–456.

Guo, S., L. Zou, Y. Liu, W. Ye, S. Cheng, S. Wang, H. Chen, D. Yin, and Y. Chang. (2021). "Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. New York, NY, USA: Association for Computing Machinery. 275–284.

Gursoy, F., R. Kennedy, and I. Kakadiaris. (2022). "A critical assessment of the algorithmic accountability act of 2022". *Available at SSRN 4193199*.

Haak, F. (2023). "Investigation of Bias in Web Search Queries". In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. Springer. 443–449.

Haak, F., B. Engelmann, C. K. Kreutz, and P. Schaer. (2024). "Investigating Bias in Political Search Query Suggestions by Relative Comparison with LLMs". In: *Companion Publication of the 16th ACM Web Science Conference*. 5–7.

Haak, F. and P. Schaer. (2021). "Perception-Aware Bias Detection for Query Suggestions". In: *Advances in Bias and Fairness in Information Retrieval: Second International Workshop on Algorithmic Bias in Search and Recommendation, BIAS 2021, Lucca, Italy, April 1, 2021, Proceedings*. Springer. 130–142.

Haak, F. and P. Schaer. (2022). "Auditing Search Query Suggestion Bias Through Recursive Algorithm Interrogation". In: *14th ACM Web Science Conference 2022*. 219–227.

Haak, F. and P. Schaer. (2023). "Qbias-A Dataset on Media Bias in Search Queries and Query Suggestions". In: *Proceedings of the 15th ACM Web Science Conference 2023*. 239–244.

Han, S., X. Wang, M. Bendersky, and M. Najork. (2020). "Learning-to-Rank with BERT in TF-Ranking". *CoRR*. abs/2004.08476. arXiv: 2004.08476. URL: https://arxiv.org/abs/2004.08476.

Hardt, M., X. Chen, X. Cheng, M. Donini, J. Gelman, S. Gollaprolu, J. He, P. Larroy, X. Liu, N. McCarthy, *et al.* (2021). "Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2974–2983.

Hardt, M., E. Price, and N. Srebro. (2016). "Equality of opportunity in supervised learning". In: *NIPS*. 3315–3323.

Hashimoto, T., M. Srivastava, H. Namkoong, and P. Liang. (2018). "Fairness Without Demographics in Repeated Loss Minimization". In: *ICML*.

Hiemstra, D. (2023). "Was Fairness in IR Discussed by Cooper and Robertson in the 1970's?" In: *ACM SIGIR Forum*. Vol. 56. No. 2. ACM New York, NY, USA. 1–5.

Hsu, B., R. Mazumder, P. Nandy, and K. Basu. (2022). "Pushing the limits of fairness impossibility: Who's the fairest of them all?" *Advances in Neural Information Processing Systems*. 35: 32749–32761.

Hu, M., H. Wu, Z. Guan, R. Zhu, D. Guo, D. Qi, and S. Li. (2024). "No Free Lunch: Retrieval-Augmented Generation Undermines Fairness in LLMs, Even for Vigilant Users". *arXiv preprint arXiv:2410.07589*.

Hu, P., L. Zhen, D. Peng, and P. Liu. (2019). "Scalable Deep Multimodal Learning for Cross-Modal Retrieval". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19*. Paris, France: Association for Computing Machinery. 635–644.

Hu, X., S. Yu, C. Xiong, Z. Liu, Z. Liu, and G. Yu. (2022). "P3 Ranker: Mitigating the Gaps between Pre-Training and Ranking Fine-Tuning with Prompt-Based Learning and Pre-Finetuning". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. New York, NY, USA: Association for Computing Machinery. 1956–1962.

Huang, Z., H. Zeng, H. Zamani, and J. Allan. (2023). "Soft Prompt Decoding for Multilingual Dense Retrieval". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. New York, NY, USA: Association for Computing Machinery. 1208–1218.

Hutchinson, B., V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. (2020). "Social Biases in NLP Models as Barriers for Persons with Disabilities". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics. 5491–5501.

Introna, L. D. and H. Nissenbaum. (2000). "Shaping the Web: Why the politics of search engines matters". *The information society*. 16(3): 169–185.

Jacobs, A. Z. and H. Wallach. (2021). “Measurement and fairness”. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

Jaenich, T., G. McDonald, and I. Ounis. (2023). “ColBERT-FairPRF: Towards Fair Pseudo-Relevance Feedback in Dense Retrieval”. In: *European Conference on Information Retrieval*. Springer. 457–465.

Jain, A., M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, and J. Baldridge. (2021). “MURAL: Multimodal, Multitask Representations Across Languages”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Ji, K., S. Pathiyan Cherumanal, J. R. Trippas, D. Hettiachchi, F. D. Salim, F. Scholer, and D. Spina. (2024). “Towards Detecting and Mitigating Cognitive Bias in Spoken Conversational Search”. In: *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction*. 1–10.

Jiang, B., Z. Tan, A. Nirmal, and H. Liu. (2023). “Disinformation Detection: An Evolving Challenge in the Age of LLMs”. *arXiv preprint arXiv:2309.15847*.

Jiang, N. and L. Li. (2016). “Doubly Robust Off-policy Value Evaluation for Reinforcement Learning”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. *Proceedings of Machine Learning Research*. New York, New York, USA: PMLR. 652–661. URL: https://proceedings.mlr.press/v48/jiang16.html.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. (2007a). “Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search”. *ACM TOIS*.

Joachims, T., A. Swaminathan, and T. Schnabel. (2017a). “Unbiased Learning-to-Rank with Biased Feedback”. In: *WSDM*.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay. (2005). “Accurately Interpreting Clickthrough Data as Implicit Feedback”. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 154–161.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. (2007b). "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search". *ACM Transactions on Information Systems (TOIS)*. 25(2): 7.

Joachims, T., A. Swaminathan, and T. Schnabel. (2017b). "Unbiased Learning-to-Rank with Biased Feedback". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 781–789.

Joachims, T., A. Swaminathan, and T. Schnabel. (2017c). "Unbiased learning-to-rank with biased feedback". In: *Proceedings of the tenth ACM international conference on web search and data mining*. 781–789.

Kairouz, P., H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.* (2021). "Advances and open problems in federated learning". *Foundations and Trends® in Machine Learning*. 14(1–2): 1–210.

Kallus, N. and A. Zhou. (2019). "The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the XAUC Metric". *NeurIPS*.

Kamiran, F. and T. Calders. (2012). "Data preprocessing techniques for classification without discrimination". *Knowledge and information systems*. 33(1): 1–33.

Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma. (2018). "Recommendation independence". In: *Conference on fairness, accountability and transparency*. PMLR. 187–201.

Kang, J. and H. Tong. (2021). "Fair graph mining". In: *CIKM*.

Karako, C. and P. Manggala. (2018). "Using image fairness representations in diversity-based re-ranking for recommendations". In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 23–28.

Kay, M., C. Matuszek, and S. A. Munson. (2015). "Unequal representation and gender stereotypes in image search results for occupations". In: *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 3819–3828.

Kearns, M., S. Neel, A. Roth, and Z. S. Wu. (2019). "An empirical study of rich subgroup fairness for machine learning". In: *Proceedings of the conference on fairness, accountability, and transparency*. 100–109.

Kearns, M., A. Roth, and Z. S. Wu. (2017). "Meritocratic fairness for cross-population selection". In: *International Conference on Machine Learning*. PMLR. 1828–1836.

Khramtsova, E., S. Zhuang, M. Baktashmotlagh, and G. Zuccon. (2024). "Leveraging LLMs for Unsupervised Dense Retriever Ranking". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24*. Washington DC, USA: Association for Computing Machinery. 1307–1317.

Kiesel, J., D. Spina, H. Wachsmuth, and B. Stein. (2021). "The meant, the said, and the understood: Conversational argument search and cognitive biases". In: *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–5.

Kilbertus, N., M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. (2017). "Avoiding discrimination through causal reasoning". In: *NIPS*. 656–666.

Kim, T. E. and F. Diaz. (2024). "Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation". *arXiv preprint arXiv:2409.11598*.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1746–1751.

Kiyohara, H., Y. Saito, T. Matsuhiro, Y. Narita, N. Shimizu, and Y. Yamamoto. (2022). "Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model". In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. WSDM '22*. Virtual Event, AZ, USA: Association for Computing Machinery. 487–497.

Kiyohara, H., M. Uehara, Y. Narita, N. Shimizu, Y. Yamamoto, and Y. Saito. (2023). "Off-Policy Evaluation of Ranking Policies under Diverse User Behavior". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23*. New York, NY, USA: Association for Computing Machinery. 1154–1163.

Kırnap, Ö., F. Diaz, A. Biega, M. Ekstrand, B. Carterette, and E. Yilmaz. (2021). "Estimation of Fair Ranking Metrics with Incomplete Judgments". In: *Proceedings of the Web Conference 2021*. 1065–1075.

Kleinberg, J. (2018). "Inherent trade-offs in algorithmic fairness". In: *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40–40.

Kleinberg, J., S. Mullainathan, and M. Raghavan. (2016). "Inherent trade-offs in the fair determination of risk scores". *arXiv preprint arXiv:1609.05807*.

Kleinberg, J., S. Mullainathan, and M. Raghavan. (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *ITCS*.

Koenecke, A., A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. (2020). "Racial disparities in automated speech recognition". *Proceedings of the National Academy of Sciences*. 117(14): 7684–7689.

Kopeinik, S., M. Mara, L. Ratz, K. Krieg, M. Schedl, and N. Rekabsaz. (2023). "Show me a" Male Nurse"! How Gender Bias is Reflected in the Query Formulation of Search Engine Users".

Krieg, K., E. Parada-Cabaleiro, G. Medicus, O. Lesota, M. Schedl, and N. Rekabsaz. (2022a). "Grep-BiasIR: a dataset for investigating gender representation-bias in information retrieval results". In: *Proceeding of the 2023 ACM SIGIR Conference On Human Information Interaction And Retrieval (CHIIR)*.

Krieg, K., E. Parada-Cabaleiro, G. Medicus, O. Lesota, M. Schedl, and N. Rekabsaz. (2023). "Grep-BiasIR: A Dataset for Investigating Gender Representation Bias in Information Retrieval Results". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. Association for Computing Machinery. 444–448.

Krieg, K., E. Parada-Cabaleiro, M. Schedl, and N. Rekabsaz. (2022b). "Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?" In: *Advances in Bias and Fairness in Information Retrieval*. Springer International Publishing. 104–116.

Kurita, K., N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. (2019). "Measuring Bias in Contextualized Word Representations". In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Ed. by M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster. Florence, Italy: Association for Computational Linguistics. 166–172. DOI: 10.18653/v1/W19-3823.

Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger. (2015). "From Word Embeddings To Document Distances". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. *Proceedings of Machine Learning Research*. Lille, France: PMLR. 957–966. URL: https://proceedings.mlr.press/v37/kusnerb15.html.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva. (2017). "Counterfactual fairness". *Advances in neural information processing systems*. 30.

Lahoti, P., A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi. (2020). "Fairness without demographics through adversarially reweighted learning". *Advances in neural information processing systems*. 33: 728–740.

Lahoti, P., K. P. Gummadi, and G. Weikum. (2019). "Operationalizing individual fairness with pairwise fair representations". *VLDB Endowment*.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *ICLR*.

Lazovich, T., L. Belli, A. Gonzales, A. Bower, U. Tantipongpipat, K. Lum, F. Huszar, and R. Chowdhury. (2022). "Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics". *Patterns*. 3(8).

Le, Q. and T. Mikolov. (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. 1188–1196.

Le Quy, T., A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. (2022). "A survey on datasets for fairness-aware machine learning". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 12(3): e1452.

Lee, N., Y. Bang, H. Lovenia, S. Cahyawijaya, W. Dai, and P. Fung. (2023). "Survey of Social Bias in Vision-Language Models". arXiv: 2309.14381 [cs.CL].

Lester, B., R. Al-Rfou, and N. Constant. (2021). "The Power of Scale for Parameter-Efficient Prompt Tuning". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3045–3059.

Li, J., D. Li, C. Xiong, and S. Hoi. (2022). "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *ICML*.

Li, L., S. Chen, J. Kleban, and A. Gupta. (2015). "Counterfactual Estimation and Optimization of Click Metrics in Search Engines: A Case Study". In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion*. Association for Computing Machinery. 929–934.

Li, L. H., M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. (2019). "VisualBERT: A Simple and Performant Baseline for Vision and Language". In: *Arxiv*.

Li, S., Y. Abbasi-Yadkori, B. Kveton, S. Muthukrishnan, V. Vinay, and Z. Wen. (2018). "Offline Evaluation of Ranking Policies with Click Models". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18*. London, United Kingdom: Association for Computing Machinery. 1685–1694.

Li, Y., H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang. (2023). "Fairness in Recommendation: Foundations, Methods, and Applications". *ACM Transactions on Intelligent Systems and Technology*. 14(5): 1–48.

Li, Y., H. Chen, S. Xu, Y. Ge, and Y. Zhang. (2021a). "Towards personalized fairness based on causal notion". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.

Li, Y., Y. Ge, and Y. Zhang. (2021b). "Tutorial on fairness of machine learning in recommender systems". In: *SIGIR*.

Lima, L., V. Furtado, E. Furtado, and V. Almeida. (2019). "Empirical analysis of bias in voice-based personal assistants". In: *Companion Proceedings of the 2019 World Wide Web Conference*. 533–538.

Liu, H. and X. Zhao. (2021). "KDD 2021 IRS Workshop". In:

Liu, H., C. Li, Q. Wu, and Y. J. Lee. (2023). "Visual Instruction Tuning". In: *Advances in Neural Information Processing Systems*. Vol. 36. 34892–34916.

Liu, T.-Y. *et al.* (2009). "Learning to rank for information retrieval". *Foundations and Trends® in Information Retrieval*. 3(3): 225–331.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. (2020). "Ro{BERT}a: A Robustly Optimized {BERT} Pretraining Approach". URL: https://openreview.net/forum?id=SyxS0T4tvS.

Lu, J., D. Batra, D. Parikh, and S. Lee. (2019). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Lu, J., G. Hernandez Abrego, J. Ma, J. Ni, and Y. Yang. (2021). "Multi-stage Training with Improved Negative Contrast for Neural Passage Retrieval". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. 6091–6103. DOI: 10.18653/v1/2021.emnlp-main.492.

Luo, D., L. Zou, Q. Ai, Z. Chen, D. Yin, and B. D. Davison. (2023). "Model-Based Unbiased Learning to Rank". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 895–903.

Ma, H., S. Guan, C. Toomey, and Y. Wu. (2022a). "Diversified subgraph query generation with group fairness". In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 686–694.

Ma, H., S. Guan, M. Wang, Y.-s. Chang, and Y. Wu. (2022b). "Subgraph query generation with fairness and diversity constraints". In: *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE. 3106–3118.

Ma, H., H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, and X. Xie. (2022c). "EI-CLIP: Entity-Aware Interventional Contrastive Learning for E-Commerce Cross-Modal Retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18051–18061.

Ma, X., X. Zhang, R. Pradeep, and J. Lin. (2023). "Zero-Shot Listwise Document Reranking with a Large Language Model". arXiv: 2305.02156 [cs.IR].

Madaio, M., L. Egede, H. Subramonyam, J. Wortman Vaughan, and H. Wallach. (2022). "Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support". *Proceedings of the ACM on Human-Computer Interaction*. 6(CSCW1): 1–26.

Makhlouf, K., S. Zhioua, and C. Palamidessi. (2020). "Survey on causal-based machine learning fairness notions". *arXiv preprint arXiv:2010.09553*.

Makhortykh, M., A. Urman, and R. Ulloa. (2021). "Detecting race and gender bias in visual representation of AI on web search engines". In: *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer. 36–50.

Mandal, A., S. Leavy, and S. Little. (2021). "Dataset diversity: measuring and mitigating geographical bias in image search and retrieval".

May, C., A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. (2019). "On Measuring Social Biases in Sentence Encoders". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics. 622–628. DOI: 10.18653/v1/N19-1063.

McInerney, J., B. Brost, P. Chandar, R. Mehrotra, and B. Carterette. (2020). "Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '20*. Virtual Event, CA, USA: Association for Computing Machinery. 1779–1788.

Meade, N., E. Poole-Dayan, and S. Reddy. (2022). "An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mehrotra, A. and N. Vishnoi. (2022). "Fair ranking with noisy protected attributes". *Advances in Neural Information Processing Systems*. 35: 31711–31725.

Mehrotra, R., A. Anderson, F. Diaz, A. Sharma, H. Wallach, and E. Yilmaz. (2017). "Auditing search engines for differential satisfaction across demographics". In: *Proceedings of the 26th international conference on World Wide Web companion*. 626–633.

Mehrotra, R., J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. (2018). "Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off Between Relevance, Fairness & Satisfaction in Recommendation Systems". In: *CIKM*. ACM.

Minka, T. and S. Robertson. (2008). "Selection Bias in the LETOR Datasets". In: URL: https://www.microsoft.com/en-us/research/publication/selection-bias-letor-datasets/.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum. (2021). "Algorithmic fairness: Choices, assumptions, and definitions". *Annual Review of Statistics and Its Application*. 8: 141–163.

Mitra, B. (2024). "Search and Society: Reimagining Information Access for Radical Futures". *arXiv preprint arXiv:2403.17901*.

Mitra, B. and N. Craswell. (2017). "Neural models for information retrieval". *arXiv preprint arXiv:1705.01509*.

Mitra, B., S. Hofstätter, H. Zamani, and N. Craswell. (2021). "Improving Transformer-Kernel Ranking Model Using Conformer and Query Term Independence". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. New York, NY, USA: Association for Computing Machinery. 1697–1702.

Morik, M., A. Singh, J. Hong, and T. 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.

Mowshowitz, A. and A. Kawaguchi. (2002). "Assessing bias in search engines". *Information Processing & Management*. 38(1): 141–156.

Nabi, R. and I. Shpitser. (2018). "Fair inference on outcomes". *AAAI*.

Nadeem, M., A. Bethke, and S. Reddy. (2021). "StereoSet: Measuring stereotypical bias in pretrained language models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics. 5356–5371. DOI: 10.18653/v1/2021.acl-long.416.

Nag, P. and Ö. N. Yalçın. (2020). "Gender stereotypes in virtual agents". In: *Proceedings of the 20th ACM International conference on intelligent virtual agents*. 1–8.

Nalisnick, E., B. Mitra, N. Craswell, and R. Caruana. (2016). "Improving Document Ranking with Dual Word Embeddings". In: *Proceedings of the 25th International Conference Companion on World Wide Web. WWW '16 Companion*. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee. 83–84.

Nangia, N., C. Vania, R. Bhalerao, and S. R. Bowman. (2020). "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics. 1953–1967. DOI: 10.18653/v1/2020.emnlp-main.154.

Narasimhan, H., A. Cotter, M. Gupta, and S. Wang. (2020). "Pairwise fairness for ranking and regression". In: *AAAI*.

Nogueira, R. and K. Cho. (2019). "Passage Re-ranking with BERT". *ArXiv*. abs/1901.04085. URL: https://api.semanticscholar.org/CorpusID:58004692.

Nogueira, R., Z. Jiang, R. Pradeep, and J. Lin. (2020). "Document Ranking with a Pretrained Sequence-to-Sequence Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics. 708–718. DOI: 10.18653/v1/2020.findings-emnlp.63.

Nogueira, R. F., W. Yang, K. Cho, and J. Lin. (2019). "Multi-Stage Document Ranking with BERT". *CoRR*. abs/1910.14424. arXiv: 1910.14424. URL: http://arxiv.org/abs/1910.14424.

Nozza, D., F. Bianchi, A. Lauscher, and D. Hovy. (2022). "Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics. 26–34. DOI: 10.18653/v1/2022.ltedi-1.4.

O'Brien, M. and M. T. Keane. (2006). "Modeling result-list searching in the World Wide Web: The role of relevance topologies and trust bias". In: *Proceedings of the 28th annual conference of the cognitive science society*. Vol. 28. 1881–1886.

Olteanu, A., F. Diaz, and G. Kazai. (2020). "When are search completion suggestions problematic?" *Proceedings of the ACM on human-computer interaction*. 4(CSCW2): 1–25.

Olteanu, A., J. Garcia-Gathright, M. de Rijke, M. D. Ekstrand, A. Roegiest, A. Lipani, A. Beutel, A. Olteanu, A. Lucic, A.-A. Stoica, *et al.* (2021). "FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval". In: *ACM SIGIR Forum*. Vol. 53. No. 2. ACM New York, NY, USA. 20–43.

Oord, A. van den, Y. Li, and O. Vinyals. (2019). "Representation Learning with Contrastive Predictive Coding". arXiv: 1807.03748 [cs.LG].

Oosterhuis, H. (2022). "Reaching the End of Unbiasedness: Uncovering Implicit Limitations of Click-Based Learning to Rank". In: *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '22*. Madrid, Spain: Association for Computing Machinery. 264–274.

Oosterhuis, H. (2023). "Doubly Robust Estimation for Correcting Position Bias in Click Feedback for Unbiased Learning to Rank". *ACM Trans. Inf. Syst.* 41(3).

Oosterhuis, H. and M. de Rijke. (2020). "Policy-Aware Unbiased Learning to Rank for Top-k Rankings". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

OpenAI. (2023). "GPT-4 Technical Report". arXiv: 2303.08774 [cs.CL].

Otterbacher, J., J. Bates, and P. Clough. (2017). "Competent men and warm women: Gender stereotypes and backlash in image search results". In: *Proceedings of the 2017 chi conference on human factors in computing systems*. 6620–6631.

Ovaisi, Z., R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva. (2020). "Correcting for Selection Bias in Learning-to-Rank Systems". In: *Proceedings of The Web Conference 2020*. 1863–1873.

Page, L., S. Brin, R. Motwani, and T. Winograd. (1998). "The pagerank citation ranking: Bring order to the web". *Tech. rep.* Technical report, stanford University.

Papakyriakopoulos, O., S. Hegelich, J. C. M. Serrano, and F. Marco. (2020). "Bias in Word Embeddings". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 446–457.

Patro, G. K., L. Porcaro, L. Mitchell, Q. Zhang, M. Zehlike, and N. Garg. (2022). "Fair ranking: a critical review, challenges, and future directions". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1929–1942.

Penha, G., E. Palumbo, M. Aziz, A. Wang, and H. Bouchard. (2023). "Improving content retrievability in search with controllable query generation". In: *Proceedings of the ACM Web Conference 2023*. 3182–3192.

Perez, E., S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. (2022). "Red Teaming Language Models with Language Models". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. 3419–3448.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by M. Walker, H. Ji, and A. Stent. New Orleans, Louisiana: Association for Computational Linguistics. 2227–2237. DOI: 10.18653/v1/N18-1202.

Pinney, C., A. Raj, A. Hanna, and M. D. Ekstrand. (2023). "Much Ado about gender: Current practices and future recommendations for appropriate gender-aware information access". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 269–279.

Pitoura, E., K. Stefanidis, and G. Koutrika. (2021). "Fairness in Rankings and Recommendations: An Overview". *The VLDB Journal*. 31(3): 431–458. DOI: 10.1007/s00778-021-00697-y.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. (2017). "On fairness and calibration". *Advances in neural information processing systems*. 30.

Pradeep, R., S. Sharifymoghaddam, and J. Lin. (2023). "Rankvicuna: Zero-shot listwise document reranking with open-source large language models". *arXiv preprint arXiv:2309.15088*.

Pradel, F., F. Haak, S.-O. Proksch, and P. Schaer. (2024). "Dynamics in Search Engine Query Suggestions for European Politicians". In: *Proceedings of the 16th ACM Web Science Conference*. 279–289.

Precup, D., R. S. Sutton, and S. P. Singh. (2000). "Eligibility Traces for Off-Policy Policy Evaluation". In: *Proceedings of the Seventeenth International Conference on Machine Learning. ICML '00*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 759–766.

Qiao, Y., C. Xiong, Z. Liu, and Z. Liu. (2019). "Understanding the Behaviors of BERT in Ranking". *CoRR*. abs/1904.07531. arXiv: 1904.07531. URL: http://arxiv.org/abs/1904.07531.

Qin, Z., R. Jagerman, K. Hui, H. Zhuang, J. Wu, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, and M. Bendersky. (2023). "Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting". arXiv: 2306.17563 [cs.IR].

Quiñonero Candela, J., Y. Wu, B. Hsu, S. Jain, J. Ramos, J. Adams, R. Hallman, and K. Basu. (2023). "Disentangling and Operationalizing AI Fairness at LinkedIn". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1213–1228.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.

Radford, A. and K. Narasimhan. (2018). "Improving Language Understanding by Generative Pre-Training".

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.* (2019). "Language models are unsupervised multitask learners". *OpenAI blog*. 1(8): 9.

Radlinski, F., P. N. Bennett, B. Carterette, and T. Joachims. (2009). "Redundancy, diversity and interdependent document relevance". In: *ACM SIGIR Forum*. Vol. 43. 46–52.

Radlinski, F., R. Kleinberg, and T. Joachims. (2008). "Learning diverse rankings with multi-armed bandits". In: *ICML*. ACM. 784–791.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *J. Mach. Learn. Res.* 21(1).

Rahmani, H. A., C. Siro, M. Aliannejadi, N. Craswell, C. L. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz. (2024). "Llm4eval: Large language model for evaluation in ir". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3040–3043.

Raj, A. and M. D. Ekstrand. (2022). “Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’22*. Madrid, Spain: Association for Computing Machinery. 726–736.

Raj, A., B. Mitra, N. Craswell, and M. D. Ekstrand. (2023). “Patterns of gender-specializing query reformulation”. *SIGIR*.

Rastegarpanah, B., K. Gummadi, and M. Crovella. (2021). “Auditing black-box prediction models for data minimization compliance”. *Advances in Neural Information Processing Systems*. 34: 20621–20632.

Ratz, L., M. Schedl, S. Kopeinik, and N. Rekabsaz. (2024). “Measuring Bias in Search Results Through Retrieval List Comparison”. In: *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V*. Glasgow, United Kingdom: Springer-Verlag. 20–34. DOI: 10.1007/978-3-031-56069-9_2.

Reimers, N. and I. Gurevych. (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

Rekabsaz, N., S. Kopeinik, and M. Schedl. (2021). “Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’21*. New York, NY, USA: Association for Computing Machinery. 306–316.

Rekabsaz, N. and M. Schedl. (2020). “Do Neural Ranking Models Intensify Gender Bias?” In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

Ren, Y., H. Tang, and S. Zhu. (2022). “Unbiased Learning to Rank with Biased Continuous Feedback”. In: *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. 1716–1725.

Richardson, M., E. Dominowska, and R. Ragno. (2007). "Predicting Clicks: Estimating the Click-through Rate for New Ads". In: *Proceedings of the 16th International Conference on World Wide Web. WWW '07*. Banff, Alberta, Canada: Association for Computing Machinery. 521–530.

Robertson, S. E. (1977). "The probability ranking principle in IR". *Journal of documentation*.

Ross, C., B. Katz, and A. Barbu. (2021). "Measuring Social Biases in Grounded Vision and Language Embeddings". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Online: Association for Computational Linguistics. 998–1008. DOI: 10.18653/v1/2021.naacl-main.78.

Roy, D., D. Ganguly, M. Mitra, and G. Jones. (2016). "Representing Documents and Queries as Sets of Word Embedded Vectors for Information Retrieval". *ArXiv*. abs/1606.07869. URL: https://api.semanticscholar.org/CorpusID:15987308.

Saito, Y. (2020a). "Asymmetric Tri-Training for Debiasing Missing-Not-At-Random Explicit Feedback". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. Virtual Event, China: Association for Computing Machinery. 309–318.

Saito, Y. (2020b). "Doubly Robust Estimator for Ranking Metrics with Post-Click Conversions". In: *Proceedings of the 14th ACM Conference on Recommender Systems. RecSys '20*. Virtual Event, Brazil: Association for Computing Machinery. 92–100.

Saito, Y. and T. Joachims. (2021). "Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances". In: *Proceedings of the 15th ACM Conference on Recommender Systems. RecSys '21*. Amsterdam, Netherlands: Association for Computing Machinery. 828–830.

Saito, Y. and T. Joachims. (2022). "Off-Policy Evaluation for Large Action Spaces via Embeddings". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR. 19089–19122.

Samar, T., M. C. Traub, J. van Ossenbruggen, L. Hardman, and A. P. de Vries. (2018). "Quantifying retrieval bias in Web archive search". *International Journal on Digital Libraries*. 19: 57–75.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". *CoRR*. abs/1910.01108. URL: http://arxiv.org/abs/1910.01108.

Sap, M., D. Card, S. Gabriel, Y. Choi, and N. A. Smith. (2019). "The Risk of Racial Bias in Hate Speech Detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics. 1668–1678. DOI: 10.18653/v1/P19-1163.

Sapiezynski, P., W. Zeng, R. E Robertson, A. Mislove, and C. Wilson. (2019). "Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists". In: *Companion Proceedings of The 2019 World Wide Web Conference. WWW '19*. San Francisco, USA: Association for Computing Machinery. 553–562.

Schick, T., S. Udupa, and H. Schütze. (2021). "Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP". *Computing Research Repository*. arXiv:2103.00453. URL: http://arxiv.org/abs/2103.00453.

Schnabel, T., A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. (2016). "Recommendations as Treatments: Debiasing Learning and Evaluation". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. ICML'16*. JMLR.org. 1670–1679.

Scholer, F., D. Kelly, W.-C. Wu, H. S. Lee, and W. 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. SIGIR '13*. Dublin, Ireland: Association for Computing Machinery. 623–632.

Schramowski, P., C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting. (2022). "Large pre-trained language models contain human-like biases of what is right and wrong to do". *Nature Machine Intelligence*. 4(3): 258–268. DOI: 10.1038/s42256-022-00458-8.

Segev, E. (2010). *Google and the digital divide: The bias of online knowledge*. Elsevier.

Sesari, E., M. Hort, and F. Sarro. (2022). "An Empirical Study on the Fairness of Pre-trained Word Embeddings". In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by C. Hardmeier, C. Basta, M. R. Costa-jussà, G. Stanovsky, and H. Gonen. Seattle, Washington: Association for Computational Linguistics. 129–144. DOI: 10.18653/v1/2022.gebnlp-1.15.

Severyn, A. and A. Moschitti. (2015). "Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 373–382.

Seyedsalehi, S., A. Bigdeli, N. Arabzadeh, B. Mitra, M. Zihayat, and E. Bagheri. (2022). "Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases". In: *Proceedings of the 25th International Conference on Extending Database Technology, EDBT 2022, Edinburgh, UK, March 29 - April 1, 2022*. OpenProceedings.org. 2:435–2:439.

Seymour, W., X. Zhan, M. Cote, and J. Such. (2023). "A Systematic Review of Ethical Concerns with Voice Assistants". *AAAI/ACM Conference on AI, Ethics, and Society*.

Shi, J.-C., Y. Yu, Q. Da, S.-Y. Chen, and A.-X. Zeng. (2019). "Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 4902–4909.

Shokouhi, M., R. White, and E. Yilmaz. (2015). "Anchoring and Adjustment in Relevance Estimation". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. Santiago, Chile: Association for Computing Machinery. 963–966.

Silva, A., P. Tambwekar, and M. Gombolay. (2021). "Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics. 2383–2389. DOI: 10.18653/v1/2021.naacl-main.189.

Singh, A. and T. Joachims. (2017). "Equality of Opportunity in Rankings". In: *NIPS Workshop on Prioritizing Online Content (WPOC)*.

Singh, A. and T. Joachims. (2018). "Fairness of exposure in rankings". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

Singh, A. and T. Joachims. (2019). "Policy learning for fairness in ranking". *Advances in Neural Information Processing Systems*. 32.

Singh, A., D. Kempe, and T. Joachims. (2021). "Fairness in ranking under uncertainty". *Advances in Neural Information Processing Systems*. 34: 11896–11908.

Steed, R., S. Panda, A. Kobren, and M. Wick. (2022). "Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics. 3524–3542. DOI: 10.18653/v1/2022.acl-long.247.

Strehl, A. L., J. Langford, L. Li, and S. M. Kakade. (2010). "Learning from Logged Implicit Exploration Data". In: *Proceedings of the 23rd International Conference on Neural Information Processing Systems. NIPS'10*. Vancouver, British Columbia, Canada: Curran Associates Inc. 2217–2225.

Su, W., X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. (2020a). "VL-BERT: Pre-training of Generic Visual-Linguistic Representations". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SygXPaEYvH.

Su, Y., M. Dimakopoulou, A. Krishnamurthy, and M. Dudik. (2020b). "Doubly robust off-policy evaluation with shrinkage". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. *Proceedings of Machine Learning Research*. PMLR. 9167–9176. URL: https://proceedings.mlr.press/v119/su20a.html.

Su, Y., L. Wang, M. Santacatterina, and T. Joachims. (2019). "CAB: Continuous Adaptive Blending for Policy Evaluation and Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. PMLR. 6005–6014. URL: https://proceedings.mlr.press/v97/su19a.html.

Sun, G., Y. Bai, X. Yang, Y. Fang, Y. Fu, and Z. Tao. (2024a). "Aligning Out-of-Distribution Web Images and Caption Semantics via Evidential Learning". In: *Proceedings of the ACM on Web Conference 2024*.

Sun, G., C. Qin, J. Wang, Z. Chen, R. Xu, and Z. Tao. (2024b). "SQ-LLaVA: Self-Questioning for Large Vision-Language Assistant". In: *The European Conference on Computer Vision (ECCV)*.

Sun, W., L. Yan, X. Ma, P. Ren, D. Yin, and Z. Ren. (2023). "Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent". *ArXiv*. abs/2304.09542.

Sutskever, I., O. Vinyals, and Q. V. Le. (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 27.

Swaminathan, A. and T. Joachims. (2015a). "Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization". *Journal of Machine Learning Research*. 16(52): 1731–1755. URL: http://jmlr.org/papers/v16/swaminathan15a.html.

Swaminathan, A. and T. Joachims. (2015b). "The Self-Normalized Estimator for Counterfactual Learning". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/39027dfad5138c9ca0c474d71db915c3-Paper.pdf.

Sweeney, L. (2013). "Discrimination in online ad delivery". *Communications of the ACM*. 56(5): 44–54.

Tam, W., X. Liu, K. Ji, L. Xue, J. Liu, T. Li, Y. Dong, and J. Tang. (2023). "Parameter-Efficient Prompt Tuning Makes Generalized and Calibrated Neural Text Retrievers". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics. 13117–13130. DOI: 10.18653/v1/2023.findings-emnlp.874.

Tan, H. and M. Bansal. (2019). "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Hong Kong, China: Association for Computational Linguistics. 5100–5111. DOI: 10.18653/v1/D19-1514.

Thomas, P., S. Spielman, N. Craswell, and B. Mitra. (2024). "Large language models can accurately predict searcher preferences". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1930–1940.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.* (2023). "Llama: Open and efficient foundation language models". *arXiv preprint arXiv:2302.13971*.

Tversky, A. and D. Kahneman. (1974). "Judgment under Uncertainty: Heuristics and Biases". *Science*. 185(4157): 1124–1131.

Tversky, A. and D. Kahneman. (1992). "Advances in Prospect Theory: Cumulative Representation of Uncertainty". *Journal of Risk and Uncertainty*. 5(4): 297–323.

Vardasbi, A., H. Oosterhuis, and M. de Rijke. (2020). "When Inverse Propensity Scoring Does Not Work: Affine Corrections for Unbiased Learning to Rank". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20*. Virtual Event, Ireland: Association for Computing Machinery. 1475–1484.

Vassimon Manela, D. de, D. Errington, T. Fisher, B. van Breugel, and P. Minervini. (2021). "Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics. 2232–2242. DOI: 10.18653/v1/2021.eacl-main.190.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Vaughan, L. and M. Thelwall. (2004). "Search engine coverage bias: evidence and possible causes". *Information processing & management*. 40(4): 693–707.

Vaughan, L. and Y. Zhang. (2007). "Equal representation by search engines? A comparison of websites across countries and domains". *Journal of computer-mediated communication*. 12(3): 888–909.

Vig, J., S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. (2020). "Investigating Gender Bias in Language Models Using Causal Mediation Analysis". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 12388–12401. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Voorhees, E. M. (2002). "The Philosophy of Information Retrieval Evaluation". In: *Evaluation of Cross-Language Information Retrieval Systems*. Ed. by C. Peters, M. Braschler, J. Gonzalo, and M. Kluck. Berlin, Heidelberg: Springer Berlin Heidelberg. 355–370.

Wang, J., Y. Liu, and X. E. Wang. (2021a). "Are gender-neutral queries really gender-neutral? mitigating gender bias in image search". *arXiv preprint arXiv:2109.05433*.

Wang, J., X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, B. Jiao, Y. Zhang, and X. Xie. (2023a). "On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective". *CoRR*. abs/2302.12095. DOI: 10.48550/arXiv.2302.12095. arXiv: 2302.12095.

Wang, L. and T. Joachims. (2021). "User fairness, item fairness, and diversity for rankings in two-sided markets". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 23–41.

Wang, N., Z. Qin, X. Wang, and H. Wang. (2021b). "Non-Clicks Mean Irrelevant? Propensity Ratio Scoring As a Correction". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21*. Virtual Event, Israel: Association for Computing Machinery. 481–489.

Wang, P., X. Mi, X. Liao, X. Wang, K. Yuan, F. Qian, and R. A. Beyah. (2018). "Game of Missuggestions: Semantic Analysis of Search-Autocomplete Manipulations." In: *NDSS*.

Wang, Y.-X., A. Agarwal, and M. Dudik. (2017). "Optimal and Adaptive Off-Policy Evaluation in Contextual Bandits". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17*. Sydney, NSW, Australia: JMLR.org. 3589–3597.

Wang, X., M. Bendersky, D. Metzler, and M. Najork. (2016). "Learning to Rank with Selection Bias in Personal Search". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 115–124.

Wang, Y., W. Ma, M. Zhang, Y. Liu, and S. Ma. (2023b). "A survey on the fairness of recommender systems". *ACM Transactions on Information Systems*. 41(3): 1–43.

Wang, Y., Z. Tao, and Y. Fang. (2022a). "A Meta-learning Approach to Fair Ranking". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2539–2544.

Wang, Y., Z. Tao, and Y. Fang. (2024a). "A Unified Meta-learning Framework for Fair Ranking with Curriculum Learning". *IEEE Transactions on Knowledge and Data Engineering*.

Wang, Y., X. Wu, H.-T. Wu, Z. Tao, and Y. Fang. (2024b). "Do Large Language Models Rank Fairly? An Empirical Study on the Fairness of LLMs as Rankers". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics. 5712–5724. DOI: 10.18653/v1/2024.naacl-long.319.

Wang, Y., P. Yin, Z. Tao, H. Venkatesan, J. Lai, Y. Fang, and P. Xiao. (2023c). "An Empirical Study of Selection Bias in Pinterest Ads Retrieval". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5174–5183.

Wang, Y., L. Lyu, and A. Anand. (2022b). "BERT Rankers Are Brittle: A Study Using Adversarial Document Perturbations". In: *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 115–120.

Wang, Z., Z. Wu, J. Zhang, N. Jain, X. Guan, and A. Koshiyama. (2024c). "Bias Amplification: Language Models as Increasingly Biased Media". *arXiv preprint arXiv:2410.15234*.

Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, *et al.* (2021). "Ethical and social risks of harm from language models". *arXiv preprint arXiv:2112.04359*.

Wilkie, C. and L. Azzopardi. (2013). "Relating retrievability, performance and length". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 937–940.

Wilkie, C. and L. Azzopardi. (2014a). "A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 81–90.

Wilkie, C. and L. Azzopardi. (2014b). "Best and fairest: An empirical analysis of retrieval system bias". In: *European Conference on Information Retrieval*. Springer. 13–25.

Wu, H., C. Ma, B. Mitra, F. Diaz, and X. Liu. (2022a). "A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation". *ACM Transactions on Information Systems*. 41(2): 1–29.

Wu, H., B. Mitra, and N. Craswell. (2024a). "Towards Group-aware Search Success". In: *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*. 123–131.

Wu, H., B. Mitra, C. Ma, F. Diaz, and X. Liu. (2022b). "Joint multisided exposure fairness for recommendation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 703–714.

Wu, X., S. Li, H.-T. Wu, Z. Tao, and Y. Fang. (2024b). "Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems". *arXiv preprint arXiv:2409.19804*.

Wu, X., Y. Wang, H. Wu, Z. Tao, and Y. Fang. (2024c). "Evaluating Fairness in Large Vision-Language Models Across Diverse Demographic Attributes and Prompts". *CoRR*. abs/2406.17974. DOI: 10.48550/ARXIV.2406.17974. arXiv: 2406.17974.

Xiao, L., Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, and M. Shaoping. (2017). "Fairness-aware group recommendation with pareto-efficiency". In: *RecSys*.

Xiong, L., C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk. (2021). "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=zeFrfgyZln.

Xu, S., D. Hou, L. Pang, J. Deng, J. Xu, H. Shen, and X. Cheng. (2023). "AI-Generated Images Introduce Invisible Relevance Bias to Text-Image Retrieval". *arXiv preprint arXiv:2311.14084*.

Yadav, H., Z. Du, and T. Joachims. (2019). "Fair Learning-to-Rank from Implicit Feedback". *CoRR*. abs/1911.08054.

Yadav, H., Z. Du, and T. Joachims. (2021). "Policy-Gradient Training of Fair and Unbiased Ranking Functions". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. New York, NY, USA: Association for Computing Machinery. 1044–1053.

Yan, L., Z. Qin, H. Zhuang, X. Wang, M. Bendersky, and M. Najork. (2022). "Revisiting Two-Tower Models for Unbiased Learning to Rank". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. New York, NY, USA: Association for Computing Machinery. 2410–2414.

Yanagi, R., R. Togo, T. Ogawa, and M. Haseyama. (2021). "Database-Adaptive Re-Ranking for Enhancing Cross-Modal Image Retrieval". In: *Proceedings of the 29th ACM International Conference on Multimedia. MM '21*. Virtual Event, China: Association for Computing Machinery. 3816–3825.

Yang, J.-H., C. Lassance, R. Sampaio De Rezende, K. Srinivasan, M. Redi, S. Clinchant, and J. Lin. (2023a). "AToMiC: An Image/Text Retrieval Test Collection to Support Multimedia Content Creation". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. New York, NY, USA: Association for Computing Machinery. 2975–2984.

Yang, K., J. R. Loftus, and J. Stoyanovich. (2020). "Causal intersectionality for fair ranking". *arXiv preprint arXiv:2006.08688*.

Yang, K. and J. Stoyanovich. (2017). "Measuring fairness in ranked outputs". *SSDBM*.

Yang, L., Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. (2023b). "Diffusion models: A comprehensive survey of methods and applications". *ACM Computing Surveys*. 56(4): 1–39.

Yang, T. and Q. Ai. (2021). "Maximizing marginal fairness for dynamic learning to rank". In: *Proceedings of the Web Conference 2021*. 137–145.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. (2019). "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Yao, L., W. Chen, and Q. Jin. (2023). "CapEnrich: Enriching Caption Semantics for Web Images via Cross-Modal Pre-Trained Knowledge". In: *Proceedings of the ACM Web Conference 2023. WWW '23*. Austin, TX, USA: Association for Computing Machinery. 2392–2401.

Yao, S. and B. Huang. (2017). "Beyond Parity: Fairness Objectives for Collaborative Filtering". In: *NeurIPS*.

Yates, A., R. Nogueira, and J. Lin. (2021). "Pretrained Transformers for Text Ranking: BERT and Beyond". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*. Ed. by G. Kondrak, K. Bontcheva, and D. Gillick. Online: Association for Computational Linguistics. 1–4. DOI: 10.18653/v1/2021.naacl-tutorials.1.

Yilmaz, E. and J. A. Aslam. (2006). "Estimating Average Precision with Incomplete and Imperfect Judgments". In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. 102–111.

Yu, L., J. Chen, A. Sinha, M. Wang, Y. Chen, T. L. Berg, and N. Zhang. (2022). "CommerceMM: Large-Scale Commerce MultiModal Representation Learning with Omni Retrieval". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4433–4442.

Yue, Y., R. Patel, and H. 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. WWW '10*. Raleigh, North Carolina, USA: Association for Computing Machinery. 1011–1018.

Zehlike, M., F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. (2017). "FA* IR: A Fair Top-k Ranking Algorithm". *CIKM*.

Zehlike, M. and C. Castillo. (2020). "Reducing disparate exposure in ranking: A learning to rank approach". In: *Proceedings of the web conference 2020*. 2849–2855.

Zehlike, M., T. Sühr, C. Castillo, and I. Kitanovski. (2020). "FairSearch: A Tool For Fairness in Ranked Search Results". In: *Companion Proceedings of the Web Conference 2020. WWW '20*. Taipei, Taiwan: Association for Computing Machinery. 172–175.

Zehlike, M., K. Yang, and J. Stoyanovich. (2022). "Fairness in ranking: A survey". *ACM Computing Surveys*. 6: 1–36.

Zellers, R., X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. (2021). "MERLOT: Multimodal Neural Script Knowledge Models". In: *Advances in Neural Information Processing Systems 34*.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. (2013). "Learning fair representations". In: *International conference on machine learning*. PMLR. 325–333.

Zerveas, G., N. Rekabsaz, D. Cohen, and C. Eickhoff. (2022a). "CODER: An efficient framework for improving retrieval through COntextual Document Embedding Reranking". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. 10626–10644. DOI: 10.18653/v1/2022.emnlp-main.727.

Zerveas, G., N. Rekabsaz, D. Cohen, and C. Eickhoff. (2022b). "Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. New York, NY, USA: Association for Computing Machinery. 2532–2538.

Zhan, J., J. Mao, Y. Liu, M. Zhang, and S. Ma. (2020). "An Analysis of BERT in Document Ranking". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. Virtual Event, China: Association for Computing Machinery. 1941–1944.

Zhang, H., R. Zhang, J. Guo, M. de Rijke, Y. Fan, and X. Cheng. (2024a). "Are Large Language Models Good at Utility Judgments?" In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24*. Washington DC, USA: Association for Computing Machinery. 1941–1951.

Zhang, J., J. Mao, Y. Liu, R. Zhang, M. Zhang, S. Ma, J. Xu, and Q. Tian. (2019). "Context-Aware Ranking by Constructing a Virtual Environment for Reinforcement Learning". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1603–1612.

Zhang, R., J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, and P. Gao. (2024b). "LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=d4UiXAHN2W.

Zhang, Y., L. Yan, Z. Qin, H. Zhuang, J. Shen, X. Wang, M. Bendersky, and M. Najork. (2023). "Towards Disentangling Relevance and Bias in Unbiased Learning to Rank". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23*. New York, NY, USA: Association for Computing Machinery. 5618–5627.

Zhao, J., T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. (2019a). "Gender Bias in Contextualized Word Embeddings". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics. 629–634. DOI: 10.18653/v1/N19-1064.

Zhao, Z., L. Hong, L. Wei, J. Chen, A. Nath, S. Andrews, A. Kumthekar, M. Sathiamoorthy, X. Yi, and E. Chi. (2019b). "Recommending What Video to Watch next: A Multitask Ranking System". In: *Proceedings of the 13th ACM Conference on Recommender Systems. RecSys '19*. Copenhagen, Denmark: Association for Computing Machinery. 43–51.

Zhou, Y., Z. Dou, Y. Zhu, and J.-R. Wen. (2021). "PSSL: Self-Supervised Learning for Personalized Search with Contrastive Sampling". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management. CIKM '21*. Virtual Event, Queensland, Australia. 2749–2758.

Zhu, Y., H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, and J.-R. Wen. (2023). "Large language models for information retrieval: A survey". *arXiv preprint arXiv:2308.07107*.

Zhuang, H., Z. Qin, K. Hui, J. Wu, L. Yan, X. Wang, and M. Bendersky. (2024). "Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics. 358–370. DOI: 10.18653/v1/2024.naacl-short.31.

Zhuang, H., Z. Qin, R. Jagerman, K. Hui, J. Ma, J. Lu, J. Ni, X. Wang, and M. Bendersky. (2023). "RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. New York, NY, USA: Association for Computing Machinery. 2308–2313.

Zhuang, H., Z. Qin, X. Wang, M. Bendersky, X. Qian, P. Hu, and D. C. Chen. (2021). "Cross-Positional Attention for Debiasing Clicks". In: *Proceedings of the Web Conference 2021. WWW '21*. Ljubljana, Slovenia: Association for Computing Machinery. 788–797.

Zou, L., H. M. andXiaokai Chu, J. Tang, W. Ye, S. Wang, and D. Yin. (2022a). "A Large Scale Search Dataset for Unbiased Learning to Rank". In: *NeurIPS 2022*.

Zou, L., C. Hao, H. Cai, S. Wang, S. Cheng, Z. Cheng, W. Ye, S. Gu, and D. Yin. (2022b). "Approximated Doubly Robust Search Relevance Estimation". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management. CIKM '22*. Atlanta, GA, USA: Association for Computing Machinery. 3756–3765.

Zou, L., S. Zhang, H. Cai, D. Ma, S. Cheng, S. Wang, D. Shi, Z. Cheng, and D. Yin. (2021). "Pre-Trained Language Model Based Ranking in Baidu Search". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4014–4022.