7-11-2022

Fairness in Information Access Systems

Michael D. Ekstrand
Boise State University

Anubrata Das
University of Texas at Austin

Robin Burke
University of Colorado

Fernando Diaz
Mila - Quebec AI Institute

Fairness in Information Access Systems

Michael D. Ekstrand[1], Anubrata Das[2], Robin Burke[3] and Fernando Diaz[4]

[1]*People and Information Research Team (PIReT), Boise State University, USA; ekstrand@acm.org*
[2]*School of Information, University of Texas at Austin, USA; anubrata@utexas.edu*
[3]*That Recommender Systems Lab, Department of Information Science, University of Colorado, USA; robin.burke@colorado.edu*
[4]*Mila - Quebec AI Institute, Canada; diazf@acm.org*

ABSTRACT

Recommendation, information retrieval, and other information access systems pose unique challenges for investigating and applying the fairness and non-discrimination concepts that have been developed for studying other machine learning systems. While fair information access shares many commonalities with fair classification, there are important differences: the multistakeholder nature of information access applications, the rank-based problem setting, the centrality of personalization in many cases, and the role of user response all complicate the problem of identifying precisely what types and operationalizations of fairness may be relevant.

In this monograph, we present a taxonomy of the various dimensions of fair information access and survey the literature to date on this new and rapidly-growing topic. We

Michael D. Ekstrand, Anubrata Das, Robin Burke and Fernando Diaz (2022), "Fairness in Information Access Systems", Foundations and Trends® in Information Retrieval: Vol. 16, No. 1-2, pp 1–177. DOI: 10.1561/1500000079.

preface this with brief introductions to information access and algorithmic fairness to facilitate the use of this work by scholars with experience in one (or neither) of these fields who wish to study their intersection. We conclude with several open problems in fair information access, along with some suggestions for how to approach research in this space.

List of Key Terms

Term	*Defined in*	*Page*
bias	1.3	9
disparate impact	3.2.3	51
disparate mistreatment	3.2.3	52
disparate treatment	3.2.3	50
fairness	1.3	9
group fairness	3.2.3	50
individual fairness	3.2.2	48
information access	1.1	7
information need	2.3	25

The index provides a more comprehensive cross-reference of terms used in this monograph.

1

Introduction

As long as humans have recorded information in durable form, they have needed tools to access it: to locate the information they seek, review it, and consume it. Digitally, tools to facilitate information access take a variety of forms, including information retrieval and recommendation systems; these tools have been powered by technologies built on various paradigms, from heuristic metrics and expert systems to deep neural networks with sophisticated rank-based objective functions. Fundamentally, these technologies take a user's *information need* (an explicit and/or implicit need for information for some purpose (Kuhlthau, 1993), such as filling in knowledge or selecting a product) and locate documents or items that are *relevant* (that is, will meet the user's need).

Throughout the history of these technologies — which we treat under the integrated banner of **information access systems** — both research and development have been concerned with a range of effects beyond a system's ability to locate individual items that are relevant to a user's information need. Research has examined the *diversity* and *novelty* of results (Santos *et al.*, 2015; Hurley and Zhang, 2011) and the *coverage* of the system, among other concerns. In recent years, this concern has extended to the *fairness* of an information access system: are

the benefits and resources it provides fairly allocated between different people or organizations it affects? Does it introduce or reproduce harms, particularly harms distributed in an unfair or unjust way? This challenge is connected to the broader set of research on fairness in sociotechnical systems generally and AI systems more particularly (Mitchell *et al.*, 2020; Barocas *et al.*, 2019), but information access systems have their own set of particular challenges and possibilities.

Fairness is not an entirely new concern for information access; various fairness problems can be connected to topics with long precedent in the information retrieval and recommender systems literature. In the context of information retrieval, Friedman and Nissenbaum (1996) and Introna and Nissenbaum (2000) recognized the potential for search engines to embed social, political, and moral values in their ranking functions. In order to assess the impact of such values, Mowshowitz and Kawaguchi (2002) developed a metric to measure a search engine's deviation from an ideal exposure of content. Although conversations often focus on bias in algorithmic ranking, Vaughan and Zhang (2007) and Vaughan and Thelwall (2004) note that bias can be introduced because of biased crawling and indexing; in particular, they describe, writing in the 2000s, how Chinese webpages were under-indexed by search engines. These observations led to discussion amongst legal scholars about the regulation of search engines (Goldman, 2005; Pasquale, 2006). Azzopardi and Vinay (2008) proposed the notion of document *retrievability* and investigated the skew in this distribution for different retrieval systems. Work on *popularity bias* (Celma and Cano, 2008; Zhao *et al.*, 2013; Cañamares and Castells, 2018) and rich-get-richer effects (Cho *et al.*, 2005), along with attempts to ensure quality and equity in *long-tail recommendations* (Ferraro, 2019), can be viewed as a type of fairness problem: the system should not inordinately favor popular, well-known, and possibly well-funded content creators. In a group recommendation, one common objective is to ensure that the various members of a group are treated fairly (Kaya *et al.*, 2020).

The work on fair information access that we present here goes beyond these problems to examine how various forms of unfairness — particularly those that arise from *social biases* (Olteanu *et al.*, 2019) — can make their way in to the data, algorithms, and outputs of informa-

tion access systems. These biases can affect many different stakeholders of an information access system; Burke (2017) distinguishes between *provider*- and *consumer*-side fairness, and other individuals or organizations affected by an information access system may have further fairness concerns.

In this monograph, we provide an introduction to fairness in information access, aiming to give students, researchers, and practitioners a starting point for understanding the problem space, the research to date, and a foundation for their further study. Fairness in information access draws heavily from the fair machine learning literature, which we summarize in Section 3; researchers and practitioners looking to study or improve the fairness of information access will do well to pay attention to a broad set of research results. For reasons of scope, we are primarily concerned here with the fairness of the information access transaction itself: providing results in response to a request encoding an information need. Fairness concerns can also arise in other aspects of the system, such as the representation and presentation of documents themselves, or in support facilities such as query suggestions (Noble, 2018). We provide brief pointers on these topics, but a detailed treatment is left for future synthesis, noting that they have not yet received as much attention in the research literature. We are also specifically concerned with fairness-related harms, and not the broader set of harms that may arise in information access such as the amplification of disinformation.

Throughout this work, we use the term **system** to describe an algorithmic system that performs some task: retrieving information, recommending items, classifying or scoring people based on their data. These systems are embedded in social contexts, operating on human-provided inputs and producing results acted upon by humans. The technical system forms one part of a broader socio-technical system.

1.1 Abstracting Information Access

Our choice to title this monograph "Fairness in *Information Access*" is quite deliberate. While there is significant technical and social overlap between information retrieval, recommender systems, and related fields, they are distinct communities with differences in terminology, problem

definitions, and evaluation practices. However, there are fundamental commonalities, and they present many of the same problems that complicate notions of fairness, including ranked outputs, personalized relevance, repeated decision-making, and multistakeholder structure. We therefore refer to them together as **information access systems** — algorithmic systems that mediate the interaction between a repository of documents or items and a user's information need.

This information access umbrella includes information retrieval, recommender systems, information filtering, and some applications of natural language processing. In Section 2, we present a fuller treatment of this integration and reviews the fundamentals of information access, both to introduce the concepts to readers who come to this paper from a general fairness background and to lay out consistent terminology for our readers from information retrieval or recommender systems backgrounds.

1.2 A Brief History of Fairness

In the pursuit of fairness in algorithmic systems and the society more generally, the authority of Aristotle's citation of Plato "treat like cases alike" is a key touchstone: a normative requirement that those who are equal before the law should receive equal treatment (Gosepath, 2011). In more recent scholarship, the study of distributive welfare extends these concepts considerably, recognizing four distinct concepts of fairness: "exogenous rights, compensation, reward, and fitness." (Moulin, 2004). *Exogenous rights*, as the term suggests, relate to external claims that a system must satisfy: equal shares in property as defined by contract, for example, or equality of political rights in democratic societies. *Compensation* recognizes that fairness may require extra consideration for parties where costs are unequal — affirmative action in hiring and college admissions are well-known examples. *Reward* justifies inequality on the basis of differing contributions: for example, increased bonuses to employees with greater contribution to the bottom line. Finally, we have *fitness*, the most nebulous category, and the one that many information access systems inhabit. The fitness principle holds that goods be distributed to those most fit to use, appreciate, or

derive benefit from them. It is an efficiency principle, where the fairest use is the one that allocates goods where the distribution achieves the maximum utility. Fitness has a natural application to information access, as we seek to locate documents and make them visible based on their utility to the user's information need.

U.S. legal theory has developed a rich tradition of anti-discrimination law, aimed at ensuring that people are not denied certain benefits (housing, work, education, financial services, etc.) on the basis of protected characteristics (race, color, religion, gender, disability, age, and in many jurisdictions, sexual orientation). It has given rise to several important concepts, such as the disparate impact standard (the idea that an allegedly discriminatory practice can be legally challenged on the grounds that it has disproportionate adverse impact on a protected group, without needing to show intent to discriminate[1]). Crenshaw (1989) points out some of the limitations of this legal framework; in particular, it has often focused on discrimination on the basis of *individual* protected characteristics, and people who have suffered harm as a result of combinations of protected characteristics (e.g. Black women being denied promotions given to both Black men and White women) have difficulty proving their case and obtaining relief. This theory of particular harms deriving from combinations of characteristics is called **intersectionality**.

Questions of fairness and discrimination have been the subject of significant discussion in many other communities as well. Educational testing, for example, has several decades of research on the fairness of various testing and assessment instruments; this history is summarized for computer scientists by Hutchinson and Mitchell (2019). Friedman and Nissenbaum (1996) provide one of the earlier examples of addressing questions of bias in computer science, pointing out how even seemingly-innocuous technical decisions may result in biased effects when a computing system is used in its social context. The last ten years have seen significant new activity on fairness in machine learning that

[1] Disparate impact is not sufficient basis to *win* a discrimination lawsuit; rather, it is the first step in a multi-stage burden-shifting framework used to decide discrimination cases under the standard. Barocas and Selbst (2016) provide an overview of the process.

forms the primary stream of algorithmic fairness research; in Section 3 we provide an introduction to this literature.

1.3 Fairness and Bias

There are many overlapping terms used to discuss issues of fairness, bias, and discrimination. While we give a fuller treatment of the vocabulary in Section 3, we will here introduce how we use these terms in this monograph. Work we cite may use them differently.

When we refer to **fairness**, we are talking about the ways a system treats people, or groups of people, in a way that is considered "unfair" by some moral, legal, or ethical standard. This is typically through effects or impacts that are not experienced in an equitable way, but can sometimes arise through the system's internal operation or representations. This definition is similar to how Friedman and Nissenbaum (1996) use the term "bias". There is not one particular definition of what constitutes fairness, as Selbst *et al.* (2019) and many others have noted; for the purpose of terminology, the important point is that we use the term to refer to normative ideas of what it means to treat people "fairly", no matter their source.

When we talk about **bias**, we are using the term in something closer to its statistical sense: we mean properties of estimators, models, measurements, and data that systematically deviate from their intended ideal target. As detailed in Section 3.1, we share an expansive view of bias with Mitchell *et al.* (2020, Section 2.2.1), noting that these biases can be the kinds of statistical biases familiar to science (systematic discrepancies between data or outputs and the underlying observable world), but they can also be societal biases in the form of systematic discrepancies between the observable world and the arguable ideal world that would arise if society eliminated all forms of illegitimate discrimination.

The key distinction in our work is that we use the term "bias" to refer to a fact of the system without making any inherently normative judgment, and "fairness" to discuss the normative aspects of the system and its effects. Some biases are themselves fairness problems; some biases cause fairness problems; some have no effect with regards to

the concerns of fairness; and some may be intentionally introduced to address a fairness problem, often by correcting for another bias. Most fairness problems arise from biases somewhere in the system, its data, or its evaluation, but we find it useful to distinguish between the technical fact and the moral, ethical, or legal concern.

1.4 Fairness and Other Responsibility Concerns

Fairness is commonly grouped together with other concerns under the banner of *responsibility* in computing systems. These concerns include:

Accountability Research on accountability examines the legal, social, and technical mechanisms by which computing systems and their operators, developers, and providers may be held accountable, usually for the human effects of their systems. This can connect directly to fairness when considering how to hold organizations accountable for ensuring their systems uphold societal goals to be fair. Such accountability can be through formal structures, such as applying anti-discrimination law to computing systems, or through informal structures such as applying pressure through publicizing the results of third-party audits.

Transparency Transparency (and its close cousin explainability) seeks to make the operation and results of algorithmic systems scrutable to users, developers, auditors, and other stakeholders so that it can be understood, reviewed, and contested. This relates to long-standing concern in information access on explanation (Tintarev and Masthoff, 2007), as well as ideas such as scrutable user models (Kay *et al.*, 2002).

Safety Information access systems can be harmful. They can distribute false information, promote fake or dangerous products, and provide support for illegal or malicious activities. These problems have received attention in the research literature, often under the general heading of *adversarial information retrieval*. See related workshops AIRWeb (Fetterly and Gyöngyi, 2009) and WebQuality (Nielek *et al.*, 2016).

Privacy Aspects of users' profiles including queries, interaction history, and usage patterns may be highly revealing of sensitive personal information: consider queries about medical symptoms or clicks on web pages for addiction counseling. It follows that information access systems have a duty to protect such information from harmful disclosure. Research on privacy-preserving recommendation seeks technical solutions to this challenge. Friedman *et al.* (2015) provide a survey of this area.

Ethics Computing ethics is concerned broadly with ensuring that the practice and products of computing adhere to appropriate ethical principles. The ACM Code of Ethics (ACM Council, 2018) specifically calls out non-discrimination, along with attention to potential harms, as an ethical obligation for computing professionals.

The report on the FACTS-IR Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval (Roegiest *et al.*, 2019) discusses how many of these concepts play out in information retrieval. In this work we are concerned with fairness, but bring in other concerns as well when they relate to fairness.

1.5 Running Examples

Throughout this monograph, we will use several examples to motivate and explain the various concepts we discuss.

Job and Candidate Search Many online platforms attempt to connect job-seekers and employment opportunities in some way. Some of these are dedicated employment-seeking platforms, while others, such as LinkedIn and Xing, are more general-purpose professional networking platforms for which job-seeking is one important component.

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

- Do users receive a fair set of job opportunities in the recommendations or ads in their feed?

- If the system assesses a match or fit score for a candidate and a job, is this score fair, or does it under- or over-estimate scores for particular candidates or groups of candidates?
- Do users have a fair opportunity to appear in search lists when recruiters are looking for candidates for a job opening (Geyik and Kenthapadi, 2018)?
- Do employers in protected groups (minority-owned businesses, for example) have their jobs fairly promoted to qualified candidates?
- What fairness concerns come from regulatory requirements?

Music Discovery The search and recommendation systems in music platforms, such as Spotify, Pandora, and BandCamp, connect listeners with artists. These discovery tools have a significant impact not only on a user's listening experience and musical enjoyment, but also on artists' financial and career prospects, due both to direct revenue from listening and the commercial and reputational effects of visibility. Some specific fairness concerns include:

- Do artists receive fair exposure in the system's search results, recommendation lists, or streamed programming?
- Does the system systematically over- or under-promote particular groups of artists or songwriters through recommendations, search results, and other discovery surfaces (Epps-Darling *et al.*, 2020)?
- Do users receive fair quality of service, or does the system systematically do a better job of modeling some users' tastes and preferences than others?
- Do recommendations reflect well a user's preferences and if not, are there systematic errors due to stereotypes of gender, ethnicity, location, or other attributes?

News News search and recommendation influences user exposure to news articles on social media, news aggregation applications, and

search engines. Such influence extends to social and political choices users might make (Kulshrestha *et al.*, 2017; Epstein and Robertson, 2015). Additionally, the filter bubble effect (Pariser, 2011; Alstyne and Brynjolfsson, 2005) may cause users to be exposed primarily to news items that reinforce their beliefs and increase polarization. Depending on the journalistic policy of the provider, news platforms may want to facilitate balanced exposure to news from across the social, political, and cultural spectrum, but this may need to be balanced with the need to de-rank malicious and low-credibility sources.

Specific fairness concerns in news discovery include:

- Does the system provide fair exposure to news on different topics or affected groups?
- Do journalists from different perspectives receive fair visibility or exposure for their content?
- Does the system reward original investigators or primarily direct readers to tertiary sources?
- Do users receive a balanced set of news content?
- Are users in different demographics or locations equally well-served by their news recommendations?

Philanthropic Giving Online platforms are increasingly a site for philanthropic giving (Goecks *et al.*, 2008), and therefore recommendation is expected to be an increasing driver of donations. Sites may take an explicitly "peer-to-peer" approach to such giving, as in the educational charity site DonorsChoose.org; this results in many possible donation opportunities for donors to select from, requiring recommendation or sophisticated search to help match donors and opportunities. As many philanthropic organizations have a social justice focus, fairness concerns are essential in developing and evaluating their information access solutions, in particular to avoid potential positive feedback loops in which a subset of causes comes to dominate results and rankings.

In philanthropic settings, we would expect fairness issues to include:

- Does the system provide fair opportunities for the various recipients / causes to have their needs supported?
- Are specific groups of recipients under- or over-represented in the recommendation results?

1.6 How to Use This Monograph

We have written this monograph with two audiences in mind:

- Researchers, engineers, and students in information retrieval, recommender systems, and related fields who are looking to understand the literature on fairness, bias, and discrimination, and how it applies to their work.
- Researchers in algorithmic fairness who are looking to understand information access systems, how existing fairness concepts do or do not apply to this set of applications, and the things that information access brings to the research space that may differ from the application settings in which fairness is usually studied.

Due to our interest in serving both of these audiences, we do not expect our readers to have significant familiarity with either information retrieval or algorithmic fairness, although some background in machine learning will be helpful. We have organized the material as follows:

- **Section 2** rehearses the fundamentals of information access systems. This will be a review for most information retrieval and recommender systems researchers; such readers should read it for the terminology we use to integrate the fields, but may wish to focus their study energy elsewhere.
- **Section 3** provides an overview of research on fairness in machine learning generally, particularly in classification. Algorithmic fairness researchers will likely find this section to be a review.
- **Section 4** lays out the problem space of fair information access, providing a multi-faceted taxonomy of the problems in evaluating and removing discrimination and related harms in such systems.

- **Sections 5 and 6** survey key literature to date (as of 2021) on fairness in information access, with pointers to research working on many of the problems identified in Section 4, focused on the two most commonly-studied stakeholders: consumers and providers (with discussion of subjects in Section 6.4).
- **Section 7** discusses the need to go beyond point-in-time views of fairness to understand fairness over time how the temporal dynamics of an information access system affect fairness.
- **Section 8** looks to future work and provides tips for research and engineering on fair information access.

Section 4 is the keystone of this work that ties the rest together; subsequent sections work out details in the form of a literature survey of several of the problems discussed in Section 4, and the preceding sections set up the background needed to understand it. For readers looking to budget their time, we recommend they ensure they have the necessary background from Sections 2 and 3, read Sections 4 and 8, and read the later sections that are relevant to their work.

1.7 Our Perspective

While we have written this monograph to be useful for researchers approaching the topic of fairness from a variety of perspectives, we think it is helpful to explicitly describe our own perspectives and motivations, as well as the position from which we approach this work and some limitations it may bring.

Information access systems need to meet a variety of objectives from multiple stakeholders. They need to deliver relevant results to their users, business value for their operators, and visibility to the creators of the documents they present; they often also need to meet a variety of other goals and constraints, such as diversity across subtopics, regulatory compliance, and reducing avoidable harm to users or society. Fairness, as we conceive of it and present it in this monograph, is not a be-all end goal, but rather another family of objectives to be considered in the design and evaluation of information access systems, and a

collection of techniques for enabling those objectives. It also does not encompass the totality of social or ethical objectives guiding a system's design. Researchers and developers need to work with experts in ethics, policy, sociology, and other relevant fields to identify relevant harms and appropriate objectives for any particular application; the concepts we discuss will be relevant to some of those harms and objectives.

We also emphasize the importance of starting with a robust *problem framing*: Section 4 is intended to help readers think about the fairness problem they are trying to solve, and position it in a landscape of information access; we have then organized our survey in Sections 5-7 around aspects of problem definition, instead of underlying techniques. Metrics and mitigations are best developed and assessed in the context of a specific, well-defined problem.

Finally, all four authors work in North America and approach the topic primarily in that legal and moral context. A Western focus, and particularly concepts of bias and discrimination rooted in United States legal theory, currently dominates thinking and research on algorithmic fairness in general. This is a limitation of the field that others have noted and critiqued (Sambasivan *et al.*, 2020); our present work acknowledges but does not correct this imbalance. While we attempt to engage with definitions and fairness objectives beyond the U.S., this work admittedly has a Western and especially U.S. focus in its treatment of the material. We look forward to seeing other scholars survey this topic from other perspectives.

1.8 Some Cautions

We hope that this monograph will help scholars from a variety of backgrounds to understand the emerging literature on fairness in information access and to advance the field in useful directions. In addition to the general concerns of careful, thoughtful science, work on fairness often engages with data and constructs that touch on fundamental aspects of human identity and experience. This work must also be done with great care and compassion to ensure that users, creators, and other stakeholders are treated with respect and dignity and to avoid various traps that result in overbroad or ungeneralizable claims.

We argue that there is nothing particularly new about this, but that thinking about the fairness of information access brings to the surface issues that should be considered in all research and development on information systems.

1.8.1 Beware Abstraction Traps

Our first caution is to beware of the allure of abstraction. Selbst *et al.* (2019) describe several specific problems that arise from excessive or inappropriate abstraction in fairness research in general. Their core argument is that the tendency in computer science to seek general, abstract forms of problems, while useful for developing tools and results that can be applied to a wide range of tasks, can cause important social aspects of technology and its impacts to be obscured.

One reason for this is that social problems that appear to be structurally similar arise from distinct (though possibly intertwined) causes and mechanisms, and may require different solutions. Sexism and anti-Black racism, for example, are both types of discrimination and fall into the "group fairness" category of algorithmic fairness, but they are not the same problem and have not been reinforced by the same sets of legal and social processes. Discrimination also varies by culture and jurisdiction, and oppression of what appears to be same group may arise from different causes and through different mechanisms in the different places in which it appears. Kohler-Hausmann (2019) argues that social constructivist frameworks for understanding group identities and experiences imply that even understanding what constitutes a group, let alone the discrimination it experiences, is inextricably linked with understanding how that group is constructed and treated in a particular society — an understanding that is inherently bound to the society in question, although there may be similarities in group construction in different contexts.

The result is that unfairness needs to be measured and addressed in each specific way in which it may appear. While general solutions for detecting and mitigating fairness-related harms may arise and be very useful, their effectiveness needs to be re-validated in context for the harms they are meant to address, a point reiterated by Dwork and Ilvento (2018).

Hoffmann (2019) similarly provides several warnings against overly simple ideas of the harms that can arise from discrimination and bias. Computational fairness inherits some of these limitations from its reference material, such as limitations of anti-discrimination law; others arise from what Hoffmann, Selbst *et al.* (2019), and others argue are reductionistic operationalizations of rich concepts. Hoffmann (2019) notes in particular—and we agree—that treating categories of personal identity as objective features in a multi-dimensional space (a natural move for computer scientists) obfuscates the role of technical and social systems in enacting and producing such categories. This move also has the effect of reducing intersectionality concerns to what can be captured by a subspace projection or similar formal operation, whether or not that corresponds to individual's lived experience.

We believe computing systems in general, and information access systems in particular, have the opportunity to *advance* the discussion of emancipation and justice, not just bring existing constructs into a new domain. Information professionals have long been concerned about issues of ethics and justice. Just as two examples, we note that Edmund Berkeley, one of the founders of the Association for Computing Machinery, was an outspoken advocate for the ethical responsibilities of computer scientists as far back as the 1960s (Longo, 2015), and the creation of Computer Professionals for Social Responsibility in the mid-1980s (Finn and DuPont, 2020). The call here is to realize that vision fully and for all people affected by information access systems.

1.8.2 Beware Limits

It is crucial to be clear about the limitations of particular fairness studies and methods. Any work will be limited, if for no other reason than the impossibility of completely solving the problem of discrimination. Those limitations should not paralyze the research community or keep researchers from doing the most they can to advance equity and justice with the resources available to them; rather, work in this space needs to be forthright and thorough about the limitations of its approach, data, and findings. Some limitations common to this space include:

- Single-dimensional attributes for which fairness is considered, when in reality people experience discrimination and oppression along multiple simultaneous dimensions.
- Binary definitions of attributes, when in reality many social dimensions have more than two categories or exist on a continuum.
- Taking attributes as fixed and exogenous, when social categories are complex and socially constructed (Hanna *et al.*, 2020).
- Incomplete, erroneous, and/or biased data (Olteanu *et al.*, 2019; Ekstrand and Kluver, 2021).

This is not to say that work on single binary attributes is not useful; research must start somewhere. But it should not *stop* there, and authors need to be clear about the relationship their work in its broader context and provide a careful accounting of its known limitations.

Some methods are so limited that we advise against their use. For example, some work on fair information access has used statistical gender recognition based on names or computer vision techniques for gender recognition based on profile pictures.[2] This source of data is error-prone, subject to systemic biases (Buolamwini and Gebru, 2018), reductionistic (Hamidi *et al.*, 2018), and fundamentally denies subjects control over their identities, so we do not consider it good practice.

1.8.3 Beware Convenience

Researchers working in this problem space also need to be careful to do the *best* research possible with available resources, and work to expand those resources to increase the quality and social fidelity of their work, and not take the path of least resistance.

One particular application pertains to this monograph itself and to its proper use and citation. It is convenient and common practice to cite survey papers to quickly summarize a topic or substantiate its

[2] We do not provide citations to support the claim that this is in use because our purpose in this paragraph is to critique a general trend, not to focus on any specific paper. Elsewhere in this monograph, we cite work making use of these techniques where it makes a relevant contribution.

relevance. While we naturally welcome citations of our work, we would prefer to be cited specifically for our contributions to the organization and synthesis of fair information access research. The purpose of much of this monograph is to point our readers to the work that others have done, and we specifically ask that you **cite those papers**, instead of — or in addition to — this one when that work is relevant to your writing and research.

2

Information Access Fundamentals

Information access refers to a class of systems that support users by retrieving items from some large repository of data. The area covers both information retrieval and recommendation systems. More concretely, the information access problem can be defined as:

> Given a repository of items and a user information need, present items to help the user satisfy that need.

The *repository* may be the results of a hypertext crawl as in web search, a catalog of products as in commercial recommendation, corporate documents as in enterprise search, or a collection of songs as in music recommendation. An *information need* refers to the latent concept or class the user seeks. Although unobserved, this need may inferred from explicit expressions from the user (e.g. a keyword query, a question, or a set of example documents) or implicit data about the user (e.g. previously consumed content, time of day). The *presentation of items* refers to the system response and might be a ranked list, a two-dimensional grid, or some other structured output. Finally, *satisfaction* is a measure of the degree to which the system response fulfilled the user's need. This may be explicit (e.g. ratings, binary feedback) or implicit (e.g. clicks, streams, purchases).

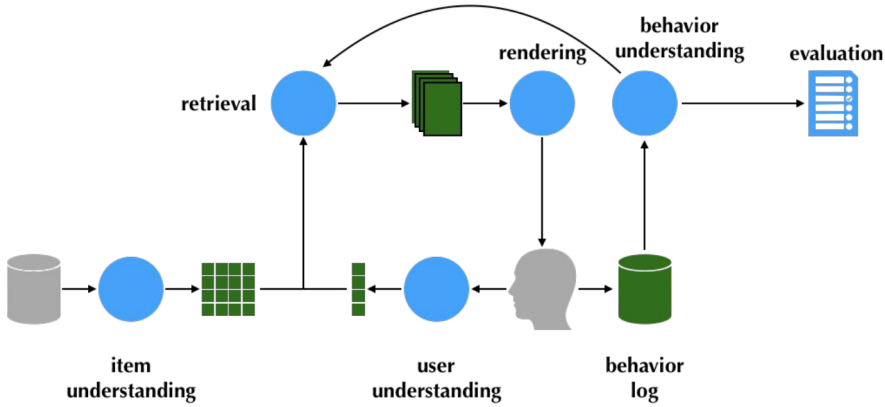

Figure 2.1: A typical information access pipeline consists of item understanding, user understanding, item retrieval, item rendering, behavior understanding and evaluation.

In this section we provide a brief overview of the fundamentals of these systems, both for our readers who are not familiar with information retrieval or recommender systems, and to provide terminology for our integration of the topics. Table 2.1 summarizes the notation we use.

2.1 System Overview

The process of meeting an information need involves several steps; Figure 2.1 shows one view of the components carrying out these steps and their relationships. These pieces include:

- Understanding (with a computationally-useful representation) the items to be retrieved, so they can be connected with information needs.
- Understanding the user and their information need, so that it can be matched to relevant items.
- Retrieving items that match the need.
- Rendering the set of retrieved items for presentation to the user.

- Understanding users' behavior, particularly in response to presentations of retrieved results, to inform future retrieval and to evaluate the system's ability to meet user needs.

Table 2.1: Summary of notation for Section 2.

$d \in \mathcal{D}$	Item in a repository
$\phi_c(d)$	Content representation of an item d
$\phi_m(d)$	Metadata representation of an item d
$\phi_u(d)$	Usage representation of an item d
$q \in \mathcal{Q}$	Information needs with the space of possible needs
$\rho_\phi(q)$	Feature-based expression of information need
$\rho_\mathcal{D}(q)$	Item-based expression of information need
$\rho_\ell(q)$	Language-based expression of information need
$\rho_{\text{global}}(q)$	Inferred global information need
$\rho_{\text{local}}(q)$	Inferred local information need
$r \in \mathbb{Z}^+$	A rank position from the set of possible ranks
$u : \mathcal{D} \to \Re$	Item utility function
$\delta : \mathbb{Z}^+ \to [0, 1]$	Rank discount function

2.2 Repository of Items

As we have defined it, information access is the process of retrieving **items** that are contained in a **repository**. Terms for this can vary; in text-oriented information retrieval, these are often called "documents" in a "corpus". The curation of this repository involves a variety of subtasks and algorithmic research, including content creation, content collection, and content representation.

Content creation refers to the complex organizational, social, economic, and political dynamics that result in the creation of an item. This includes items that may be created in relative isolation by a hobbyist (e.g. a self-produced song) or as a result of more complex coordination amongst many contributors (e.g. a major motion picture); together, we call these the **providers** of the item. Regardless of this variation in apparent scale, such cultural objects are always an artifact of social relationships (Becker, 1982). As such, each item reflects potentially a major stakeholder or stakeholders upon whom a livelihood may rest but also a network of supporting stakeholders, each of whom may have a variety of incentives for participating in the item's creation and consumption.

Content collection refers to the processes and policies involved in adding or removing items from the repository. In the simplest case, content collection involves receiving a static archive of items or documents from some external party. More often, the repository is dynamic, provided as a feed of items to add or remove from the repository, as with news articles or products. However, it is often the case that the repository designer has some control over content collection, either because the content must be found (as in hypertext crawling (Pandey and Olston, 2008)), curated (e.g. by removing low quality or adversarial content, or simply caching for performance reasons), or contracted for. Without loss of generality, we refer to the repository of items at the time when the user engages with the information access system as $\mathcal{D}$, a set of indices mapping into the items.

Content representation refers to the information we have about an individual item. **Item representation** in general consists of three parts: content, metadata, and usage data. In most cases, the content of an item is static and is an artifact created by the author(s) at some fixed point in time. This might be the text of an academic article, the pixels of an image, or the audio file associated with a song. We refer to the content representation of an item $d \in \mathcal{D}$ as $\phi_c(d)$.

The metadata of an item expresses information about the content, including when it was authored, the name of the author, the genre of the item, etc. Metadata may also be inferred by a machine such as with "learned representations" or algorithmically-assigned genre labels. Furthermore, metadata may be dynamic, changing as a function of the world around the item. For example, this might include frequency of access (outside of a search or recommendation context) or popularity of the author. We refer to the metadata representation of an item $d \in \mathcal{D}$ as $\phi_m(d)$.

Finally, usage data about an item refers to the historic interaction between information needs (or their expression) and the item. This might include, for example, the set of users who consumed the item, the queries for which this item was clicked, etc. Usage features can be seen as a special type of metadata that have three unique properties: they are often more biased by system decisions; they are often highly suggestive of relevance; and they are updated over time as users interact

with the system and its items. We refer to the usage data representation of an item $d \in \mathcal{D}$ as $\phi_u(d)$.

2.3 Users and Information Needs

The system is used by **users** (or *consumers*), who rely on it to meet some **information need**. When a user approaches the system, they may do so for a variety of reasons with varying degrees of specificity and explicitness. Their need may be isolated (e.g. answering a specific question), a function of mood (e.g. playing a genre of music), or an element of a task the user is involved in (e.g. finding reviews before making a purchase). Whatever the need, we represent the space of needs as $\mathcal{Q}$, and individual needs as $q \in \mathcal{Q}$.

An information need can arrive in a variety of ways. In most information retrieval systems, users can explicitly express their need. One such family of expressions are "feature-based" expressions, where the searcher explicitly describes item representation values likely to be found in relevant items. For example, a keyword **query** suggests that the items containing the keywords may be more relevant than those not; a faceted query may indicate a time range or class of items more likely to be relevant. We represent the feature-based expression of an information need $q \in \mathcal{Q}$ as $\rho_\phi(q)$. Alternatively, the user may provide a set of example relevant or non-relevant items (Smucker and Allan, 2006). We represent the item-based expression of an information need $q \in \mathcal{Q}$ as $\rho_\mathcal{D}(q)$. Finally, the user may express their need by some other means such as a natural language question or description.[1] We represent this expression of an information need $q \in \mathcal{Q}$ as $\rho_\ell(q)$.

An information need may also be expressed implicitly, and these implicit aspects can be global or local. An implicit *global* expression reflects relatively stationary properties of the users across access sessions. This may include demographic information about the user, their previously accessed items, or information provided or inferred about

[1] Although, because both natural language questions and text items share a symbolic representation, it is tempting to treat a natural language question as a feature-based expression, the generating processes behind questions and documents are sufficiently different that this would be a mistake.

their preferences. We represent the global implicit expression of an information need $q \in \mathcal{Q}$ as $\rho_{\text{global}}(q)$, which serves as the **user representation**. An implicit *local* expression reflects properties of the user or their interactions that typically vary across access sessions. This may include the items viewed or interacted with within the session. Importantly, this also includes the 'surface' of the search or recommendation platform itself, since the type of access or need may be suggested by the user's entry into the system; for example, a "discovery" feature versus a "mood" feature in a music streaming platform. Implicit local aspects of the need also includes contextual information such as time and location, which are meaningful parts of the description of many needs. We represent the local implicit expression of an information need $q \in \mathcal{Q}$ as $\rho_{\text{local}}(q)$. We note that the distinction between local and global can be fluid, especially in hierarchical information access (e.g. a collection of sessions with a shared goal) (Jones and Klinkner, 2008).

2.4 Presentation

Given an information need and our repository, there are various ways to present **results** to users within the constraints and capabilities of the particular user interface. In this section, we describe several of the types of presentation formats that influence algorithm and system design.

A single **turn** refers to a user approaching the system with a one-shot request and receiving an isolated system response. A recommendation system home page or isolated search query are examples of this type of presentation context. The simplest presentation involves the system providing a single item to satisfy the user's information need. This might occur in limited surfaces like small screens or audio interfaces. In situations where the interface allows, a ranked list of items can be presented, which the user can serially scan from the top down; much historical work on information retrieval and recommender systems assumes such a layout, and it underpins common evaluation metrics. A popular way to present image-based results is a two-dimensional grid layout (Guo *et al.*, 2020; Xie *et al.*, 2019). Finally, immersive environments allow for three-dimensional layouts of items (Leuski and Allan, 2000).

When items have text summaries (e.g. a description or teaser content), they can be presented alongside the item identifier (e.g. a title or URL) to let the user inspect the item without consuming it in its entirety (e.g. reading a whole web page, watching a whole movie). In some cases, the system can detect a specific part of the content to help the user better make a decision (Luhn, 1960; Tombros and Sanderson, 1998; Metzler and Kanungo, 2008).

2.4.1 Interaction and Sessions

In reality, almost every information access system involves multiple interactions in order to satisfy a single information need. Users engage with search systems in **sessions** composed of multiple queries. Users engage with recommendation systems by navigating through pages, websites, and application interfaces before settling on relevant content. These settings consist of system decisions (such as those in single turn scenarios) interleaved with implicit and explicit user feedback, allowing the algorithm to adapt its behavior. For example, streaming audio systems (e.g. radio-like interfaces) involve a sequence of single item decisions interleaved with user skips or other behavior. A dialog-based recommendation system similarly exhibits multiple turns before resolution. At a temporally-extended level, an information need or task may be spread across multiple sessions, such as assembling a bibliography for a class or survey.

2.4.2 Rankings

No matter the final presentation mode, systems typically operate in terms of **rankings** of items. The simplest, and historically most common, ranking is to sort items by decreasing relevance according to the probability ranking principle (Robertson, 1977). Items can be ranked in other ways as well (see Section 2.6.5); the system may display the items in order in which the underlying algorithm ranked them, or it may rearrange them (e.g. into a grid) or combine the results of multiple rankings (e.g. the rows of rankings display common in video streaming services), but the ranking is the fundamental unit of algorithmic output considered in most information access research, and constitutes

the 'decisions' that a search or recommendation algorithm typically makes.

2.5 Evaluation

Classic information access **evaluation** of a system decision uses a model of the user interaction together with an item-level utility estimate (e.g. an item label or behavioral signal) to measure the extent to which the system has **satisfied** the information need (Chandar *et al.*, 2020).

There are two forms in which this evaluation can take place: situated and simulated. *Situated evaluation* places the algorithm or system in front of real users, operating the system in the exact environment in which it will be deployed. *Simulated evaluation* uses data and algorithms to create a controlled, repeatable simulation of user behavior and metrics. Offline evaluation, including off-policy evaluation, is an example of this approach. As with any simulation, the assumption is that the simulator — or log data — can be used in a way that predicts performance in a situated setting (Ferro *et al.*, 2018). Situated evaluation, on the other hand, while considering the state of the world as potentially non-stationary and ephemeral, is more costly in terms of risk to users, time, and engineering effort.

2.5.1 Estimating Item Utility

Given some expression of an information need and a repository, we would like to estimate the **utility** of each item to the information need *for evaluation purposes*. We contrast this from utility estimation performed by a system *before* presenting results to users because, in the evaluation context, we can exploit information about the information need and item that may be unavailable at decision time. This information may be unavailable because of prohibitive labeling cost (e.g. manually rating documents in response to search query) or because the user has yet to observe or experience the item.

Explicit labels refer to utility estimates manually provided by some person. User-based explicit labels can be gathered by providing users with the option to rate items during or after the access or as part of

some on-boarding process, as is performed in many recommendation systems. It is important to understand the user's annotation mindset, especially when gathering labels during an access task. Are these labels judgments about the quality "in the context" or "in general"? There is often ambiguity in the precise semantics of a rating of label that can cause confusion in evaluation. In some situations, primarily in search, the information need and items are easy to understand by a third party and, therefore, "non-user" assessors can be employed to gather item quality estimates. Editorial explicit labels can be gathered using explicit guidelines, reducing some of the labeling ambiguity found in user-based explicit labels. However, these labels are limited by the interpretability of the information need and relationship of items to it. In many recommendation contexts, we expect utilities estimates to be user-specific and perhaps idiosyncratic expressions of interest and taste to which a third-party evaluator would not have access.

Implicit labels refer to utility estimates derived from observed user behavior believed to be suggestive of utility. Logged signals like clicks, streams, purchases, and bookmarking all have been found to have this property in information retrieval (Kelly and Teevan, 2003). These signals depend critically on the domain. A click may be suggestive of utility in web search but not in news recommendation where the headline alone might satisfy the user's information need. The precise relationship between post-presentation behavioral signals is often complex deserving its own modeling effort.

Feedback like clicks, streams, or bookmarking all are meant to capture the *instantaneously-measurable* utility of an item and may not provide an accurate reflection of its longer-term utility to a particular task or mood, or the lifetime value of the product to the user. Clicking or inspecting a document may be suggestive of the utility of an item within the context of an individual ranking or system ordering but may have low utility for a user's higher level goal or task. As such, when there is measurable, longer-term utility (e.g. task completion, item purchase, user retention), we can attempt to attribute that longer-term utility to earlier system decisions using methods from causal inference, reinforcement learning, and multicriteria optimization (Mann *et al.*, 2019; Chandar *et al.*, 2020).

2.5.2 Evaluating System Decisions

Although understanding item utility is critical to evaluating an information access system, these items are presented in a structured output, usually a ranked list. So, while the previous section described how we might estimate an item's utility, we are really interested in measuring the system's ability to provide users with high utility items in the course of their interactions with the system in support of their long term goal.

Situated Evaluation

In an online environment, we can adopt situated evaluation by inspecting interaction data such as short-term and longer-term behavioral signals (e.g. clicks, streams, or purchases). This data can be used to estimate the utility of consumed items (Section 2.5.1) and, by aggregating across users and needs, measure system performance. In practice, because we are often comparing pairs of systems, these metrics are computed in well-designed A/B tests (Kohavi *et al.*, 2020).

Simulated Evaluation

In offline evaluation, we can simulate online evaluation using a combination of data and user **browsing models**. Simulation allows for highly efficient testing of algorithms, and avoids any risk to users, who may not be interested in being subject to A/B experiments. If simulation data and models are freely shared amongst the research community, this allows for standard benchmarking protocols, such as have been adopted in NIST's TREC evaluations. While the use of labeled data and an evaluation metric are not often considered simulation, one reason we adopt this framing is that it centers the key objective that should guide design decisions in such an evaluation: credibly estimating likely performance with respect to the information-seeking task(s) the system is designed to support.

The data involved in offline evaluation consists of estimated item utility for a set of information needs (Section 2.5.1), often called "qrels" (for query relevance). Traditional evaluation such as used in most TREC competitions primarily uses explicit labels from assessors.

Given this data and a system decision (e.g. a ranking), the offline evaluation model simulates the behavior of a user engaging with the system decision. At the termination of the simulation, a metric value (or values) is returned. Although many offline evaluation methods can be computed analytically, each encodes a specific information access model, carrying assumptions about user behavior.

In general, many analytic evaluation metrics involve an inner product between the vector of item utilities at each rank and a rank-discount factor (Carterette, 2011). This can be represented in general form as:

$$\mu(\pi) = \sum_{r=1}^{|\mathcal{D}|} \delta(r) u(\pi_r) \tag{2.1}$$

where $\delta : \mathbb{Z}^+ \rightarrow [0, 1]$ is the rank **discount function** mapping from possible ranks to the $[0, 1]$ interval, $\pi \in S_{|\mathcal{D}|}$ is the system ranking, and $u : \mathcal{D} \rightarrow \Re$ is a item utility (for this information need). The implicit user browsing model is encoded in δ, which reflects the probability estimate that a user will inspect an item at a particular rank r. So, for binary relevance, precision-at-k can be defined with:

$$\delta_{\text{P@}k}(r) = \begin{cases} \frac{1}{k} & r \leqslant k \\ 0 & r > k \end{cases} \tag{2.2}$$

This corresponds to the expected utility for a user that randomly selects an item from amongst the top k ranks. For rank-biased precision (Moffat and Zobel, 2008),

$$\delta_{\text{RBP}}(r) = (1 - \gamma)\gamma^{r-1} \tag{2.3}$$

where γ is a hyperparameter. This browsing model models a user who sees an item at rank position r with exponentially decreasing probability. This is proportional to the expected utility of the last item inspected.

2.5.3 Evaluation Encodes Human Values

Because many modern information access systems optimize performance toward an evaluation metric or utility, system designers should be aware of the variety of personal and societal values imbued in those definitions.

Guidelines for human assessors, heuristics to detect utility in log data, and selection of longer-term utility all are contestable phenomenon and should be interrogated before being incorporated into a production pipeline or presumed to be objective in an academic paper (Stray, 2020).

2.6 Algorithmic Foundations

Information access algorithms are responsible for locating the items that will satisfy the user's information need. These algorithms often work by estimate the utility of the document to an information need through a **scoring function** $s(q, d)$, and using these utility estimates to rank results. These rankings may also be stochastic (Diaz *et al.*, 2020), expressed as a **policy** $\boldsymbol{\pi}(q)$ defining a distribution over (possibly truncated) rankings of documents. Deterministic rankings can be treated as a policy placing all probability mass on a single ranking.

Our purpose here is not to provide a comprehensive treatment of information access algorithms but to provide enough information on their key concepts that scholars familiar with machine learning can understand the particularities of information access that are important for understanding how its fairness concerns differ from those of other ML applications. Readers can find details on algorithmic foundations in a variety of information retrieval textbooks (Rijsbergen, 1979; Manning *et al.*, 2008; Croft *et al.*, 2010).

There are several distinguishing factors between different algorithmic approaches to meeting information needs:

- What data about needs and items is used, and how is it represented?
- Is utility directly estimated or learned through optimization?
- For what objective are utility estimates optimized?
- How are utility estimates used to produce the final ranking?

For example, many techniques make use of item content $\phi_c(d)$ in some way, but the family of recommendation algorithms called **collaborative filters** ignore document content entirely and use patterns

in historical user-item interaction records from $\rho_{\text{global}}(d)$ as the sole basis for estimating relevance and ranking items.

2.6.1 Vector Space Model

A long-standing approach to representing items with substantial content information, especially documents and queries is to represent them as vectors in a high-dimensional space. In the *bag of words* model, this is done by treating each word (or *term* $t \in \mathcal{T}$) as a dimension and locating documents in $|\mathcal{T}|$-dimensional space based on the relative frequency with which they use different terms (Salton *et al.*, 1975). With such an approach, a document d's ϕ_c and a need q's ρ_ϕ (and/or ρ_ℓ) can be represented as vectors $\mathbf{x}_d$ and $\mathbf{x}_q$, and the system can estimate relevance by comparing the vectors (often using the cosine $s(d|q) = \cos(\mathbf{x}_d, \mathbf{x}_q)$).

The precise definition of the vectors is usually based on the *term frequency*, and — for document representations — normalized by the *document frequency* to give greater weight to terms appearing in fewer documents (and thus more useful for locating documents relevant to information needs for which those terms appear in the query). One common representation is *term frequency — inverse document frequency*, or TF-IDF:

$$x_{d,t} = \text{TF}(d,t) \cdot \text{IDF}(d)$$

These vectors form the rows of the $|\mathcal{D}| \times |\mathcal{T}|$ *document-term matrix* $\mathbf{X}$. If both document and query vectors are normalized to unit vectors, then the similarity can be estimated with the inner product $s(d,q) = \mathbf{x}_d \cdot \mathbf{x}_q$.

Vector space models can also be used without query or document content. Pure collaborative filtering algorithms compute recommendations based solely on users' ρ_{global} by using a *ratings matrix* — a partially-observed $|\mathcal{U}| \times |\mathcal{D}|$ matrix recording users' past interactions with items, either through implicit feedback (whether or how frequently the user has consumed the item) or explicit feedback (an ordinal or real-valued preference, such as a rating on a 5-star scale). Relevance estimation is often done via neighborhoods: finding other users similar to the user needing information and scoring items by a weighted average of these neighbors' ratings (Herlocker *et al.*, 2002), or finding items similar

to those rated by the current user (Deshpande and Karypis, 2004). In both cases, ρ_{global} consists of the user's past ratings or interactions with items.

One important similarity between the document-term matrix and the ratings matrix is that they are both *sparse* and *incomplete*. Most documents do not contain most words (sparsity); most documents do not contain all synonyms and paraphrases (incompleteness). Most users have not consumed most items (sparsity); most users have not provided ratings for all items they have consumed (incompleteness).

The term-based vector space model can also be integrated with user history for *content-based recommendations*; in this case, a transformation of the items in the user's history, such as the sum or average of their term vectors, is used as a query vector to locate items that match ρ_{global} on the basis of their associated features or terms instead of user-item interaction patterns.

2.6.2 Embedding and Optimizing Utility

Two significant developments move beyond the vector space model, and form a key component of modern information access algorithms. The first is representing (or *embedding*) documents and information needs with a common lower-dimensional space (called a *latent feature space*), resulting in an **item embedding**. Introduced for information retrieval as *latent semantic analysis* (Deerwester *et al.*, 1990), one approach is to take the truncated singular value decomposition of the document-term matrix $\mathbf{X} = \mathbf{D}\Sigma\mathbf{T}$. The left and right singular matrices of this decomposition provide a low-rank representation of documents ($\mathbf{D} \in \mathbb{R}^{|\mathcal{D}| \times k}$) and a map from term vectors into this vector space ($\mathbf{T} \in \mathbb{R}^{k \times |\mathcal{T}|}$). This facilitates compact and efficient document representation and comparison (for example, similar documents can be located by comparing their vectors in $\mathbf{D}$), and allows documents to match information needs that do not contain any of their terms but do contain synonyms.

Variants of this technique have seen widespread use in recommender systems (Koren *et al.*, 2009), where the ratings matrix is decomposed into low-rank representations of users and items. The sparsely-observed nature of the ratings matrix and the computational complexity of SVD

have led to a space of approximation techniques for matrix factorization. In practice, the user-feature and item-feature matrices are inferred through stochastic gradient descent (Funk, 2006) or alternating least squares (Takács and Tikk, 2012) to minimize the reconstruction error on the observed values of the ratings matrix.

Learning decompositions through optimization is an instance of the second of these developments: estimating utility through machine learning models. This is now the basis of most current information access research. Models can become quite complex and incorporate multiple aspects of items and information needs simultaneously, but their fundamental operation is to learn a function $s(d|q)$, that estimates the item's relevance to the given need q based on observations, such as search result clicks, purchases, or product ratings. These estimates can be rating predictions, in the case of recommender systems with explicit ratings, or other estimates such as the probability of the user clicking a result (studied under the label of *CTR prediction*) or the probability that the document is relevant to the user's need (a common framing for search). Learned utility models can also be implemented directly on a vector-space representation, as in the SLIM technique for learning neighborhood interpolation weights (Ning and Karypis, 2011).

2.6.3 User Modeling

The **user model** is the component of a personalized information access system — recommendation, personalized search, and other systems that respond to ρ_{global} containing a user's historical interaction with the system — that represents the user's preferences for use in the rest of the system. It is often latent in the system architecture; for example, in the vector space model of collaborative filtering, the user model is just the set or bag of items with which the user has interacted with in the past, and in an embedding-based system it is the user's embedding (and the means by which this embedding is computed).

One early user model for book recommendation by Rich (1979) represented users as probability distributions over a set of *stereotypes* that was incrementally refined through text-based dialogue with the user. Contemporary systems often learn latent user models from the

user's history, typically taking the form of a **user embedding** (Section 2.6.2). Since user activity occurs over time and users' preferences are not necessarily stable, some techniques such as that of Koren (2010) decompose user preference into long-term *stable* preference and short-term *local* preference, to separately model the user's persistent taste and ephemeral current interests.

The fundamental idea, common to all these techniques and many more, is that a personalized system will often have a representation of ρ_{global}, computed by a suitable statistical model, that is then used by the final scoring and ranking logic in order to estimate the relevance of an item to the user's information need in accordance with their personal preferences. Whether this is an entirely separate model, computing and storing user model representations to be used as input data for a ranker, or a sub-component of an end-to-end ranking model depends on the system architecture.

2.6.4 Learning to Rank

Learned relevance models are not limited to learning pointwise relevance estimates given observed utility signals. *Learning to rank* (Liu, 2007) moves past learning to predict pointwise item-need utility and instead optimizes the learning model to rank items consistently with their ability to meet the user's information need. In the case of binary relevance (the simplest form of utility label), the system learns to rank relevant items above irrelevant items. Such systems often still learn a scoring function $s(d|q)$, but the function is optimized to produce scores that correctly order documents, regardless of the precise values of those scores, instead of its ability to estimate relevance judgments.

One approach is to optimize *pairwise ranking loss*; a key example of this in recommender systems is Bayesian Personalized Ranking (BPR; Rendle *et al.*, 2009). Given an information need q, BPR optimizes the probability $\mathrm{P}(s(d_+|q) > s(d_-|q))$ for a randomly-selected relevant item d_+ and irrelevant item d_- by maximizing $\text{logistic}\left(s\left(d_+|q\right) - s\left(d_-|q\right)\right)$ (this is equivalent to maximizing the area under the ROC curve). Pairwise loss functions can be applied to a wide range of ranking models (including matrix factorization, neighborhood models, and neural networks) for both recommendation and information retrieval.

2.6.5 Re-ranking

Many information access techniques do not depend only on a single ranking step. **Re-ranking** approaches use a base ranking model — which can be directly estimated, learned from pointwise optimizations, or a learning-to-rank model — and adjust the outputs to achieve additional goals.

One application of re-ranking is to improve the **diversity** of results. **Maximum marginal relevance** (MMR) adjusts the ranking to balance, at each position, maximizing $s(d|q)$ with minimizing the similarity between the new item and previous items (Carbonell and Goldstein, 1998). The idea behind this approach is that relevance is not independent: if one item does not meet the user's need, then another very similar item is also unlikely to meet their need, and therefore the second position should go to an item that is likely to match the query given that the first document did not (Goffman, 1964). Ziegler *et al.* (2005) provides another approach to diversifying a result list that operates purely over item orderings instead of balancing similarity or relevance scores. As we will see later in this survey, re-ranking is a common approach for improving certain types of fairness in recommendation lists.

Another use for re-ranking is to improve the efficiency of a system facilitating access to a large repository. Learned relevance or ranking models often have significantly higher computational cost than vector space models, which can be heavily optimized through index structures. One approach, therefore, is to use a simple first-pass ranker to retrieve a pool of candidate items that is significantly larger than the final result list but much smaller than the repository. A more complex ranking algorithm, possibly employing modern deep learning models, re-ranks these candidate items to produce the final ranking.

3

Fairness Fundamentals

As noted in Section 1.2, the second decade of the 21st century has brought significant attention to the issue of fairness in computing systems, particularly (but not exclusively) machine learning and statistical tools for use in decision support contexts (Mitchell *et al.*, 2020). This arises at the intersection of increasing adoption of machine learning technologies in sectors with direct real-world effects such as healthcare, public policy, and law enforcement, and extensive discussions on justice and equity in society at large. Fairness but one of several dimensions in which the social impact of computing systems are under scrutiny; a community has coalesced around studying "Fairness, Accountability, and Transparency"[1], and the topic arises in discussion fora on AI ethics and in the various communities working directly on artificial intelligence, machine learning, data mining, and information retrieval, among others.

In this section, we provide a brief overview of the landscape of algorithmic fairness: its fundamental concepts and definitions, sources of unfairness or bias, some methods for reducing the unfairness of machine learning systems, and pointers to additional research topics. We refer readers to papers by Mitchell *et al.* (2020) and Barocas and

[1]The FAccT conference (https://facctconference.org) and related venues.

Selbst (2016), as well as the in-progress book by Barocas *et al.* (2019), for more in-depth treatment of these topics.

Algorithmic fairness in general is concerned with going beyond the aggregate accuracy or effectiveness of a system — often, but not always, a machine learning application — to studying the distribution of its positive or negative effects on its subjects (distributional harms) and the ways those subjects are represented by and in the system (representational harms; Crawford, 2017). Much of this work has been focused on fairness in classification or scoring systems; Mitchell *et al.* (2020) provide a catalog of the key concepts in this space, and Hutchinson and Mitchell (2019) situate them in the broader history of fairness in educational testing where many similar ideas were previously developed. There are many ways to break down the various concepts of algorithmic fairness that have been studied in the existing literature, which we later summarize in Table 3.2.

While "algorithmic fairness" is the label used for this research, it does not encapsulate one goal, but rather covers a spectrum of equity concerns. Selbst *et al.* (2019) identify a number of "abstraction traps" surrounding fairness research, one of which is the *formalism trap*:

> Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms.

This fundamental contextuality and contestability, combined with the incompatibility of disparate notions of fairness (Friedler *et al.*, 2021), imply that universal fairness is not an achievable (or arguably even meaningful) concept.

What can be done, and what much of the research on algorithmic fairness is concerned with, is to identify specific ways in which a system may be *unfair*, and develop tools to measure and mitigate this unfairness. Determining the ways in which a system may be unfair, and how those failure cases should be assessed and addressed in any particular application, is a domain- and application-specific process that needs to be carried out in consultation with a broad set of stakeholders and subject experts.

Table 3.1: Summary of notation for Section 3.

$v_i \sim V$	Input covariates for instance i
$y_i \sim Y$	Observed outcome for instance i
$\psi(v_i)$	Computed score for instance i
$\delta(v_i)$	Decision for instance i
a_i	Sensitive attributes for instance i
x_i	Non-sensitive attributes for instance i
$\Delta_{\text{in}}(v_i, v_j)$	Distance between observations of instances i and j
$\Delta_{\text{dec}}(\delta(v_i), \delta(v_j))$	Distance between decisions for instances i and j

There are many different terms that are used, sometimes interchangeably and sometimes with nuanced differences, in scholarship and surrounding discourse on algorithmic fairness. The terms "bias" (Baeza-Yates, 2018), "fairness" (Dwork *et al.*, 2012), "discrimination" (Kamiran and Calders, 2012), "equity" (Katell *et al.*, 2020), and "justice" (Lee *et al.*, 2019), among others, are used variously by different authors to discuss the interrelated concerns and phenomena under study. In much of the literature, terms and concepts are borrowed from legal analysis, particularly the scholarly tradition around U.S. anti-discrimination law (Barocas and Selbst, 2016); Hoffmann (2019) discusses some of the limitations of this trend. D'Ignazio and Klein (2020, p. 60) call for challenging the power structures of data science and its applications, and identify some terms (including "bias" and "fairness") with perspectives that uphold existing power structures and others (such as "equity" and "justice") with questioning and dismantling those structures. Not all authors use terms in the same way, or with the same nuances.

Just as there are many ways in which a system can be unfair, there are also many places in which unfairness can be introduced in any particular system. In Section 3.1, we provide an overview of how each step of the pipeline may introduce biases that give rise to unfairness.

In this section, we focus primarily on supervised classification problems and adopt the notation and conventions of Mitchell *et al.* (2020), summarized in Table 3.1. Given an individual i with observed features v_i and observed outcome y_i, modeled as samples from random variables V and Y, the goal is to learn and evaluate a decision function $\delta(v_i)$. δ is often structured as a (possibly probabilistic) decision process f applied to an underlying score ψ, such that $\delta(v_i) = f(\psi(v_i))$. The goal

is to make "correct" decisions, so that $y_i = \delta(v_i)$; we may also wish to accurately estimate probabilities, such that $\psi(v_i) = \mathrm{P}(Y = 1|V = v_i)$.

Throughout this section, we will use two examples: lending and hiring. Lending is a classic setting for classification: the decision function $\delta(v_i)$ determines whether i will be granted or denied a loan, and the goal is to grant loans to all applicants who will pay them (we assume for the moment no limit on funds to lend). A hiring decision is closely related to the task of job and candidate recommendation introduced in section. A decision might be to choose whether or not to move on a candidate to the next level of review, but the score $\psi(v_i)$ could also be used to rank candidates for presentation as recommendations to hiring manager.

One key concept in fairness, that we discuss in more detail in Sections 3.2.2 and 3.2.3, is the idea of *individual* and *group* fairness. Individual fairness is concerned with treating similar individuals similarly, while group fairness is concerned with identifying and addressing differences between groups of data subjects. These groups are often the kinds of groups treated in anti-discrimination law, such as gender, race, or religion; to represent these groups, we can decompose an individual's covariates $v_i = (a_i, x_i)$, where a_i is the **sensitive attribute(s)** (sometimes called protected characteristics) recording group association and x_i is the other attributes (non-sensitive attributes).

3.1 Sources of Unfairness

Before discussing how to measure and precisely define unfairness, we first discuss where it can enter into the system. The harms that are discussed under the rubric of fairness typically arise from some kind of bias, where observations or outcomes are different — at least in expectation — from what they would be if the bias or unfairness were not present. These biases can arise in many places: in society, in the observations that form our data, and in the construction, evaluation, and application of decision support models (Suresh and Guttag, 2019).

Friedler *et al.* (2021) use the idea of *construct spaces* and *observation spaces* to define biases in terms of skews in observation and decision-

making processes. The **construct feature space** (CFS) contains the 'true' features that we would use to make decisions in an ideal system, such as the applicant's ability to repay a loan or the job candidate's ability to carry out the duties of a position.

The CFS is unobservable; instead, we have access to the **observation feature space** (OFS), which is the result of an observation process that results in the input features for the actual decision process (v_v); the OFS also contains the observed outcomes y_i for training data. For example, in looking at a loan application, the decision system would see the applicant's current salary, assets, and other financial details and these would form the set of observations on which the decision is made. Similarly, a hiring system might see the candidate's recent employment history, educational credentials, etc. as its OFS.

Decisions are made on the basis of these features, yielding the **observation decision space** (ODS); the goal is for these decisions to match what they would be if made on accurate and perfect information (without discrimination), the **construct decision space** (CDS).

We focus our own discussion on sources of unfairness on stages of the pipeline or feedback loop in algorithmic decision support. Figure 3.1 shows the various elements of this process, from initial observations to human response the model that affects the natural and social world that will be observed in future iterations; it is the ML equivalent of Figure 2.1. Unfairness can arise at any stage in this process, and may be propagated, mitigated, exacerbated, or even re-introduced either at that stage or at further downstream stages.

First, the **world itself** may be unfair or unjust. One source of such injustice is historical and ongoing discrimination. For example, redlining in United States housing policies (Rothstein, 2017) prevented Black Americans from purchasing homes in wealthier neighborhoods; the neighborhoods where they were allowed to live typically had lower investment in parks, schools, and other amenities to improve quality of life and childhood development. An entirely accurate survey of family wealth disaggregated by race will reveal significant racial disparities, not because there are innate racial differences in the ability to develop generational wealth (a difference in Friedler *et al.*'s *construct space*

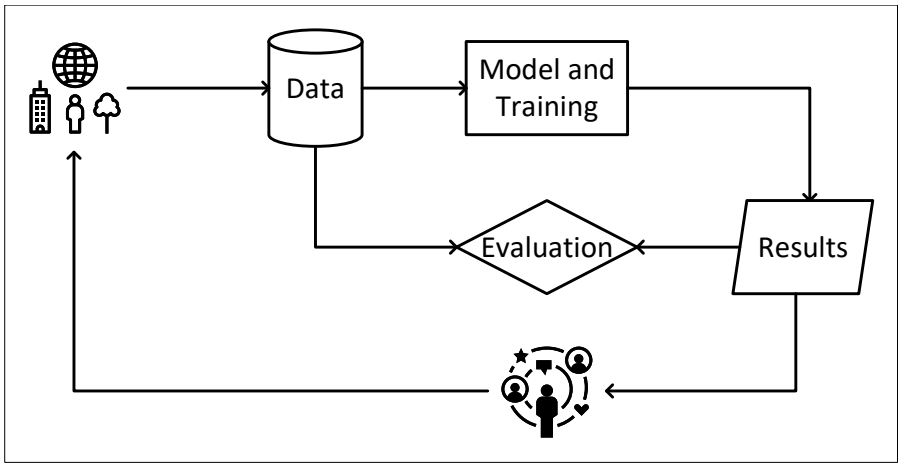

Figure 3.1: A view of the machine learning pipeline. The environment (natural and social) is observed through data. This data is used to develop and train a model, which produces results of some form, and is used to evaluate those results. The results are also acted on by people, individually or socially, who impact the environment for the next round of modeling.

(2021)), but because Black residents were prohibited by private-sector and governmental policy from accessing the same opportunities for wealth-building through home ownership as their White would-be neighbors. Thus, the OFS in a loan decision system, focused on financial indicators, would inevitably look quite different for the average Black applicant as opposed to the average White one.

Redlining has further follow-on effects that result from things such as disparate quality of education. The effects of such discrimination also often extend beyond the end of official practice and policy. As with lending, a hiring decision maker looking at educational aspects of candidates' applications may see differences in attainment (OFS) reflective of this history rather than being reflective of differences in ability (CFS) that are relevant for the position.

Group size can also play a role in some types of unfairness (Rolf *et al.*, 2021). If one group is smaller than another, naïve modeling may be more accurate at inferring and predicting things for the majority group, and may be more likely to be incorrect for minority groups.

In addition to capturing bias and injustice in the world, **data collection** may introduce unfairness into the system. Sampling strategies determine who is considered for inclusion in the data. Response or submission bias, where some people are more likely to volunteer information or respond to requests than others, can also introduce additional unfairness. Selecting and defining variables is crucial as observations or proxies need to be valid and unbiased measurements of the target construct (to avoid *measurement bias*), and their response needs to be consistent across different groups in a population (*measurement invariance*, Steenkamp and Baumgartner, 1998). For example, it wasn't until September 2021 that US mortgage lender Fannie Mae opted to use on-time rental payments as a measure of creditworthiness, in addition to traditional loan repayment data (Lerner, 2021). Collecting only loan repayment data excludes individuals who have historically had less need to borrow or who have religious objections to lending. The selection of particular data as a marker of creditworthiness therefore excludes evidence that might be favorable to a borrower.

The problems above can all introduce bias in the data under its own terms: assuming a set of information goals, the data is biased with respect to its ability to meet those goals. However, data collection can also be biased by the perspectives that inform what is being measured and how, and how the relevant constructs are defined ("When you measure include the measurer."; Hammer, 2021). The decisions that are made in collecting — from defining its initial goals to defining the variables to record — reflect the perspectives of the people involved in the process; without broad stakeholder engagement, the data set may be biased *in concept* relative to the needs and goals of a subset of the people it will affect, and without clear documentation of the perspectives and assumptions that went into its design, these biases may be undetectable (Hutchinson *et al.*, 2021). For example, it is well-known that in schools in the US, Black and White students are subject to disciplinary actions at very different rates (Okonofua and Eberhardt, 2015), and this is at least partly due to the actions of children and youth being perceived differently by teachers and administrators, conditioned by race (Okonofua *et al.*, 2016).

Many variables also require a codebook to determine how the variable is recorded, particularly if it is categorical in nature, and these codebooks reflect specific perspectives in how observations should be recorded. This becomes particularly salient with sensitive personal attributes such as gender or race; racial categories are often adapted from administrative data collection efforts, but these categorizations vary across time and space and are the results of substantial political processes in determining how to record a fundamentally social construct (Hanna *et al.*, 2020). Finally, data needs to be collected and interpreted in light of its social and cultural context to avoid unfairly disregarding local knowledge and perspectives.

Mitchell *et al.* (2020, Section 2.1) provide a statistically-oriented view on the sources we have discussed so far: **societal bias** results in deviations between the "world as it should and could be" and the "world as it is", and **statistical bias** results in "systematic mismatch" between the world as it is and the collected data and observations, including both sampling biases and measurement error. Redlining is an example of societal bias, because it results in people living in different housing situations than they would if there was no racial discrimination in housing policy (the world as it should be); a systematic mismatch between actual housing situations and their records in the data, for example due to differential reporting, would be a statistical bias.

Machine learning **models** can also introduce unfairness. Such unfairness may arise from direct use of sensitive attributes (e.g., gender and race) in the model. Models may learn to discriminate indirectly from other proxy variables present in the data. The objective functions for which a model is optimized further encode specific perspectives about what constitutes a "good" model, sharing the respective challenges in determining how to define, collect, and encode data.

Unfairness can arise in **evaluating** a machine learning algorithm or application in various ways. All of the problems we have discussed regarding input data apply to evaluation, to the extent that the data is being used to evaluate the model (and it usually is, at least as an initial validation step, even when the final evaluation will involve the results of application deployment). Further, perspectives captured in the definition of success can skews evaluation outcomes, as success for

some stakeholders does not mean success for all (Barocas *et al.*, 2021). It is crucial to identify who are the stakeholders of the system, and determine whose utility is reflected in the evaluation metric(s). This connects to additional questions on the accountability of the system: Who gets to make the decision on who is being served through the design choices? To whom are they accountable?

For example, if a loan decision model is evaluated on the basis of historical data from lenders, it will most likely have thorough information about those individuals to whom loans were given, but little or no information about those whose loans were denied. Thus, the system can learn about false positive decisions but not false negatives, magnifying whatever blind spot caused such errors in the first place (O'Neil, 2017).

Details in evaluation can also make a difference in the system's fairness (Barocas *et al.*, 2021). Measuring performance averaged over all subjects will prioritize performance on the majority group, while disaggregating and reweighting performance metrics can favor systems that perform equally well across groups regardless of group size.

Finally, unfairness may be introduced by **human response** to the system's output: humans and algorithms do not necessarily compose (Srivastava *et al.*, 2019). Model output might influence stakeholders to respond differently, and their response may in some cases be inversely correlated with computational fairness. For example, Green and Chen (2019) found that providing algorithmic risk scores *increased* racial disparity in human assessments of risk in a laboratory setting, even if the scores themselves were racially fair; Albright (2019) found similar results in a study of actual judges' decision-making behavior. Other social factors may skew human response to a system; for example, community support in assisting loan application and repayment might skew individual's response to loan prediction model outcomes.

As we note, unfairness may enter the system at any of these points, and each requires different interventions and measurements. Further, bias that is removed at one stage may be re-introduced in another, such as a model re-learning bias from a proxy variable even though the data set removed disparate representation, or a human responding in a biased fashion to unbiased predictions or scores.

3.2 Problems and Concepts

Measuring and mitigating unfairness in a system requires us to identify several things:

1. Who is experiencing unfairness?
2. How does that unfairness manifest?
3. How is that unfairness determined or measured?

The answers to each of these questions flow from normative principles that motivate why and how a particular fairness study or intervention is being undertaken. Being clear about these principles and related assumptions and goals is crucial to doing coherent work, evaluating its ability to meet its goals, and assessing the relevance and appropriateness of those goals to the social problem in question.

As we noted in the introduction of this section, there has been a shift in the field's discourse from pursuing fairness as a potentially-universal goal in itself to the perspective we describe of identifying and addressing specific fairness-related harms. These harms and approaches for addressing them can be categorized according to a number of concepts, summarized in Table 3.2. As with our notation, we draw heavily from the work of Mitchell *et al.* (2020) in framing this section, although our organization is somewhat different.

Table 3.2: Summary of concepts in algorithmic fairness and harms.

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

3.2.1 Distribution and Representation

The first axis we consider is *harms of distribution* vs. *harms of representation* (Crawford, 2017). **Distributional harms** arise when someone is denied a resource or benefit (Crawford, 2017); unfairly denying loans, for example, to a group of people.

Representational harms arise when the system represents groups or individuals incorrectly, either in its internal representation (e.g. word embeddings that encode stereotyped expectations in the latent embedding space (Bolukbasi *et al.*, 2016)) or in how it represents those people to others. Systems can cause direct representational harm, by misrepresenting people to themselves or others, and representational harms can also cause distributional harms by affecting how the system makes decisions and allocates resources.

Most literature so far on algorithmic fairness considers distributional harms; representational fairness is often introduced as a tool for reducing distributional harms.

3.2.2 Individual Fairness

Another axis of unfairness often considered is *individual* versus *group* fairness. **Individual fairness** sets the goal that similar individuals should be treated similarly: given a function that can measure the similarity of two individuals with respect to a task, such as the ability of job applicants to perform the duties of a job, individuals with comparable ability should receive comparable decisions (Dwork *et al.*, 2012). This fairness concept is grounded in the normative principle that like cases should be treated alike (Binns, 2020), a notion of justice that traces back to Aristotle.

Individual fairness is typically operationalized with a task-specific distance metric Δ_{in} between individuals and another Δ_{dec} between decisions (or decision distributions, in the case of probabilistic decision processes). Formally, the decision process is deemed fair if similar individuals receive similar decisions:

$$\Delta_{\text{in}}(v_i, v_j) \leqslant \delta \implies \Delta_{\text{dec}}(\delta(v_i), \delta(v_j)) \leqslant \epsilon$$

There are several important things to note about about this metric:

- While it constrains decisions on similar instances, it makes no requirements on dissimilar instances: making the same decision for highly dissimilar individuals does not violate individual fairness.
- It depends on the existence of a fair distance function Δ_{in}; if the distance function is unfair in some way (e.g. job candidates of the same skill but different races are further apart than candidates with the same race but differing skill levels), individual fairness cannot correct for that.
- It effectively requires the decision process $\delta(v_i)$ is probabilistic; in fact, individual fairness is impossible to fully satisfy with non-probabilistic discrete decisions (Friedler *et al.*, 2016).

The dependence on a fair distance function is particularly important to consider when identifying the assumptions that underly applications of individual fairness. Friedler *et al.* (2021) identify axiomatic assumptions for providing different kinds of fairness; one, "what you see is what you get" (**WYSIWYG**), is the assumption that available data (the OFS of v_i) is an unbiased representation of underlying reality; that is, there is no systemic bias or discrimination that affects the data gathering process. Under the WYSIWYG assumption, individual fairness is directly applicable: observations v_i can be compared with an applicable distance function, and this suffices to enable the use of the individual fairness constraint to achieve the goal of treating like instances alike.

However, if systemic discrimination is present, either due to a gap between the world's ideal and current states or biases in the data gathering process (Mitchell *et al.*, 2020), then we cannot assume that similarity in the OFS is equivalent to similarity in the CFS, and thus individual fairness is only treating individuals similarly if they are similar in the (biased) observation space; it makes no guarantees about their treatment with respect to their similarity in unobserved construct space. Friedler *et al.* (2021) characterized one family of commonly-used assumptions to address such discrepancies as "we're all equal" (**WAE**): taking as an axiom the idea that different groups are fundamentally the

same with respect to the task, and thus any systematic discrepancy in observation space (e.g. members of different racial groups tending to be dissimilar in the OFS) is the result of discrimination and should be corrected for; Friedler *et al.* (2021, p. 140) note that the assumption can be interpreted either as "members of different groups are the same" or "members of different groups should be treated the same for the purposes of our task", and the resulting math is equivalent. The WAE view is seldom taken in individual fairness literature, but Binns (2020) argues, and Friedler *et al.* (2021, p. 142) concur, that this is not a fundamental limitation of individual fairness, as WAE can be used to construct a distance function that takes group-based discrimination into account, for example by performing group-wise normalization of features.

3.2.3 Group Fairness

Group fairness is concerned with ensuring that different **groups** have comparable experiences with the system in some way. The groups in question are often gender, race, ethnicity, religion, and other group associations used in anti-discrimination law, but the goals and definitions of group fairness are not limited to these groups. As noted previously, group membership is usually formally denoted through a sensitive attribute a_i. Often two groups are considered: a **protected group** $a_i = \blacktriangledown$ and a **dominant group** (sometimes called the *unprotected group* or *majority group*) $a_i = \blacktriangle$, and the goal of the system is to ensure that the protected group is not unfairly (mis)treated.

As with individual fairness, group fairness can flow from either WYSIWYG or WAE assumptions; it is also possible to conceive of WAE as a prior instead of an axiom, but this is seldom done explicitly in the relevant literature.

There are many ways to categorize group fairness constructs; in this section, we are going to organize them along the lines of legal concepts that inspired them.

Disparate Treatment

Disparate treatment is when members of different groups are intentionally treated differently (Barocas and Selbst, 2016, Section II.A).

The straightforward way for disparate treatment to manifest is as a direct property of the model, and can be removed by simply not using group membership as an input to the model.

Disparate treatment can also arise through the modelling and feature engineering processes, by selecting features known to correlate with group membership for the purpose of treating groups differently. When such intention exists, it makes more sense to address it at a process and policy level; its statistical effects will be equivalent to disparate impact.

Disparate Impact and WAE

Disparate impact is when different groups have different impact from the system's decisions: they experience decisions at different rates. In algorithmic fairness, this is formalized through **statistical parity** measures. Given a sensitive attribute a, statistical parity is satisfied if decisions are independent of group membership:

$$\mathrm{P}(\delta(v) = 1|a) = \mathrm{P}(\delta(v) = 1)$$

In the two-group case, this is often defined as parity in outcomes between the two groups:

$$\mathrm{P}(\delta(v) = 1|a = \blacktriangledown) = \mathrm{P}(\delta(v) = 1|a_i = \blacktriangle)$$

Statistical parity measures thus reflect a "we're all equal" assumption (Friedler *et al.*, 2021); if different groups are fundamentally the same in their loan qualification, then we expect them to repay loans at the same rate, and thus they should receive positive decisions at the same rate.

In U.S. anti-discrimination law, the disparate impact doctrine (Barocas and Selbst, 2016) is a test for potential discrimination in regulated decision-making processes such as employment and housing. It is usually operationalized via the "four-fifths rule"; a decision-making process δ, such as a part of the hiring procedure like a strength test, violates disparate impact under this standard if $\mathrm{P}(\delta(v) = 1|a = \blacktriangledown) < 0.8\mathrm{P}(\delta(v) = 1|a = \blacktriangle)$ — that is, the protected-group pass rate is less than four-fifths of the dominant-group pass rate.

Crucially, however, disparate impact is only one part of a broader scheme for determining whether unlawful discrimination has occurred.

A finding of unlawful discrimination based on disparate impact requires that (1) disparate impact occurs, and (2) it either does not serve a legitimate business purpose, or there would be a less discriminatory way of achieving that business purpose. This is implemented through a burden-shifting framework:

1. The plaintiff shows the challenged practice has disparate impact.
2. The defendant shows a legitimate business purpose for the practice.
3. The plaintiff shows a less discriminatory mechanism that would achieve the business purpose.

There is much subtlety in how these rules are implemented and the burden of proof at each stage needed in order for the defendent to be liable for violating anti-discrimination law; Barocas and Selbst (2016) provide more detail. The key point for our purposes in this monograph, however, is that statistical parity measures are useful for detecting where discrimination may be occurring, and they are useful objectives in situations where "WAE" is the appropriate normative assumption, but they are often not sufficient evidence of discrimination, particularly for establishing liability.

Error-Based Constructs and WYSIWYG

The next family of group fairness constructs is based on classification or prediction error. Crucially, these metrics make a WYSIWYG ("what you see is what you get") assumption (Friedler *et al.*, 2021), at least for the outcome variables y_i: they assume recorded outcomes are correct and unbiased, and the goal is to use these as a reference point for ensuring that groups are not mistreated. Methods optimizing these objectives may vary in their assumptions about the bias in v_i.

Error parity, sometimes called **disparate mistreatment**, ensures different groups do not experience erroneous decisions at different rates, conditioned on their true outcomes (Zafar *et al.*, 2017). In our lending example, if $\text{FNR}_\blacktriangledown > \text{FNR}_\blacktriangle$, then the protected group is more likely to be denied loans that they would pay off. Fair FNR can be operationalized through an independence constraint:

$$\mathrm{P}(\delta(v_i) = 0 | y_i = 1, a_i) = \mathrm{P}(\delta(v_i) = 0 | y_i = 1)$$

Similar constraints can be derived for other metrics such as FPR.

Recall parity, sometimes called **equality of opportunity**, ensures that members of different groups are equally likely to receive a favorable positive decision conditioned on positive outcome (Hardt *et al.*, 2016). Under this objective, the decision process is fair when:

$$\mathrm{P}(\delta(v_i) = 1 | y_i = 1, a_i) = \mathrm{P}(\delta(v_i) = 1 | y_i = 1)$$

A third category of group fairness objectives that rely on outcomes look at the predictive utility of the decision process. This takes a couple of flavors; we can look at **predictive value parity** in the decision process, and require that decisions for each group have the same positive predictive value (Chouldechova, 2017):

$$\mathrm{P}(y_i = 1 | \delta(v_i) = 1, a) = \mathrm{P}(y_i = 1 | \delta(v_i) = 1)$$

We can also define similar constructs on any marginal of the confusion matrix (Mitchell *et al.*, 2020); the metrics in this section are not an exhaustive list.

We can also look at **calibration parity**, requiring that the underlying scores are equally well-calibrated for each group (Kleinberg *et al.*, 2017):

$$\mathrm{P}(y_i = 1 | \psi(v_i), a) = \mathrm{P}(y_i = 1 | \psi(v_i))$$

Finally, in settings where the system is learning stochastic decision policies, such as reinforcement learning and multi-armed bandit scenarios, Joseph *et al.* (2016) and Joseph *et al.* (2018) have proposed **meritocratic fairness**, which prohibits the system from preferring a less-qualified candidate over a more-qualified one: if y_i is a continuous measure of qualification and $y_i \geqslant y_j$, then $\mathrm{P}(\delta(v_i) = 1) \geqslant \mathrm{P}(\delta(v_j) = 1)$. This construct prevents preference inversions, but does not place any bound on how large the gap in decision probabilities can be; the system can strongly prefer a mildly more qualified candidate without violating meritocratic fairness.

In practice, any of these metrics or objectives knowing the outcome y at the time of decision making. They are useful in training and

evaluating supervised classification algorithms under the WYSIWYG assumption for the historical training labels, however.

Error-based metrics have intuitive appeal, because they encode a notion of fairness that, on its face, seems quite desirable: that if two people are both qualified for a beneficial decision, their race, gender, or other group membership should not affect the decision process. They are similar in that respect to individual fairness, using recorded outcomes as the definition of "similar". There are, however, at least three important limitations for the use of these metrics:

- WYSIWYG is a strong assumption about the accuracy and lack of bias in training labels; in some cases, this assumption amounts to assuming away the problem we are trying to solve.
- There are fundamental tradeoffs between them. The Chouldechova-Kleinberg theorem (Chouldechova, 2017; Kleinberg *et al.*, 2017) states that it is impossible to simultaneously equalize more than two different error parity metrics unless the underlying base rates are equal or the classifier is perfect. Equal FPR, equal FNR, and equal PPV are all desirable properties, but in the presence of unequal base rates and imperfect models, they cannot be simultaneously achieved. Pleiss *et al.* (2017) document incompatibilities between calibration and error-based parity, particularly when base rates are not equal. Friedler *et al.* (2021), however, note that the WAE assumption amounts to assuming the base rates *are* equal, so this set of tradeoffs is no longer in effect; this assumption also often entails biased error in the training and evaluation labels.
- Observed outcomes, for either training or evaluation, are only available from a subset of the data (Ensign *et al.*, 2018). In our lending example, repayment data is only available for loans that have been issued — if the bank does not make a loan, they cannot observe if it is repaid. This is very connected to the dynamic in information access that we only obtain feedback on results that are shown to users.

3.2.4 Awareness, Treatment, and Impact

An important early result in algorithmic fairness (Dwork *et al.*, 2012) is that "fairness through unawareness" — that is, trying to achieve fairness by completely ignoring protected group status — does not work. Group identity often correlates with other variables which can serve as proxies for group membership (Feldman *et al.*, 2015); for example, different education or income levels between groups due to societal discrimination. Ignoring protected group membership can therefore result in models that are unfair under any of a number of definitions. Individual fairness addresses this through the notion of *similarity with respect to task* (Dwork *et al.*, 2012), which may need to compensate for group differences (Binns, 2020). Statistical parity addresses this through a WAE assumption, resulting in metrics that require groups to experience positive decisions at similar rates. Error parity addresses this through enforcing group parity in decision (in)accuracy, which often requires group labels at least in the evaluation — if not the training — stage of model-building.

Lipton *et al.* (2018) address the general question of whether disparate treatment — explicitly treating members of different groups differently — is necessary in order to reduce disparate impact, and argue that disparate treatment is more effective and easier to reason about than more indirect fairness interventions aimed at reducing disparate impact.

3.2.5 Motivations and Theories

In addition to different assumptions, and related to different specific goals, fairness objectives can also flow from different fundamental philosophies about the purpose and function of fairness. In U.S. legal theory, there are — broadly speaking — two different motivations for anti-discrimination law (Barocas and Selbst, 2016; Xiang and Raji, 2019). Under **anti-classification**, the goal of anti-discrimination is to remove the protected characteristic from the decision process: race is not a factor in whether or not someone gets a job or a loan. The theory of **anti-subordination**, however, says that this does not go far enough, because the effects of past discrimination carry forward in time. Under this theory, anti-discrimination law and practice need to actively work

to reverse the effects of past discrimination and oppression. Both of these theories can be found motivating fairness literature; sometimes they will converge to some of the same technical constructs, at least as an intermediate step, but they lead to different end goals and sometimes different rhetoric. The distinction in rhetoric we discussed at the beginning of Section 3 is related to this distinction in policy motivations.

3.3 Mitigation Methods

Just as bias can enter the system at different stages in the pipeline (Figure 3.1), mitigation techniques can also be applied at different stages. In this section we briefly outline some approaches from the existing machine learning literature; more details can be found in fair machine learning survey papers (Mehrabi *et al.*, 2019; Caton and Haas, 2020).

Great care is needed in selecting and evaluating sites of intervention for improving the equity of a system. Different types and sources of fairness may be best addressed by interventions at different points, but not necessarily at the source. Further, we cannot assume that fairness composes (e.g. improving the fairness in one stage does not guarantee that downstream stages in the decision-making process do not re-introduce unfairness or introduce new kinds of unfairness); Dwork and Ilvento (2018) argue it is necessary to assess the fairness of the entire system in the context of its actual use and application.

3.3.1 Preprocessing

One intervention site is to remove biases from the input data. Simply removing sensitive attributes does not necessarily lead to a non-discriminatory model outcome, as from other features in the dataset might encode information for inferring the sensitive attribute (Dwork *et al.*, 2012; Feldman *et al.*, 2015).

Kamiran and Calders (2012) propose four different methods for de-biasing data:

- Suppressing sensitive attributes or attributes correlated to the sensitive attributes; this can reduce discrimination in downstream tasks in some cases.

- "Massaging" the data by altering class labels from negative to positive for sensitive groups and vice versa until discrimination is minimized; this extends another technique by Kamiran and Calders (2009).
- Re-weighting the data by carefully assigning weights to certain inputs to reduce discrimination.
- Stratified sampling strategies to repeat or skip samples to reduce discrimination selectively.

Feldman *et al.* (2015) propose a repair procedure for data sets by altering observed features (x) in the data set to eliminate their utility for predicting sensitive attributes (a). This is a purely correlational or predictive approach, that removes any correlation between sensitive and insensitive attributes regardless of potential causal connections.

Salimi *et al.* (2019) introduce the idea of *interventional fairness*, using causal directed acyclic graphs to represent functional interactions between variables; they then divide features into *admissible* and *inadmissible*, where admissible features are those that justifiably influence a decision outcome. They then define fairness as when all outcomes are causally independent for any combination in the superset of inadmissible variables, and require classifiers to be trained on data sets that satisfy this notion of conditional independence. They finally repair the database by inserting and deleting certain tuples to ensure it satisfies the conditional independence constraint.

Beyond bias arising from the initial data set, feedback loops can feed model bias back into the data set. For example, in predictive policing, the only feedback the system receive are influenced by the decision already made by the system, such as which neighborhood to patrol, and crime reported in those neighborhoods only. These feedback loops can result in substantially increased discrimination. Ensign *et al.* (2018) document this phenomenon and provide a mitigation strategy that selectively filters the feedback on the model decisions.

3.3.2 Representation Learning

Many machine learning models operate by learning representations that can be used for (possibly multiple) downstream tasks. Imposing fairness constraints on the representations may lead to non-discriminatory output in the downstream tasks (Zemel *et al.*, 2013; Madras *et al.*, 2018; Lahoti *et al.*, 2019). The key idea in fair representation learning is to, as Zemel *et al.* put it, "lose any information that can identify whether the person belongs to the protected subgroup, while retaining as much other information as possible".

Representational learning maps the data distribution to a latent distribution where the latent distribution satisfies some desired properties. Fair representation learning can be formulated as multi-objective optimization problem, simultaneously minimizing information loss and removing information related to sensitive attributes. Several approaches include probabilistic mapping of the input data to prototypes (Zemel *et al.*, 2013), matrix transformation (Lahoti *et al.*, 2019), and fairness as adversarial objectives (Madras *et al.*, 2018; Feng *et al.*, 2019; Beutel *et al.*, 2019)

3.3.3 Fairness-Aware Decision Models

We can also alter the objectives of the decision model itself to include fairness. This typically takes the form of incorporating one or more of the objectives described in Section 3.2 into the model's objectives or training process; they therefore face the tradeoffs and incompatibilities inherent to the various constructs and metrics. One common approach ins **regularization**: incorporating one or more fairness objectives penalty terms to the loss function to discourage unfair models. In many cases, existing loss functions are enhanced with regularization terms in order to strike a balance between non-discrimination and accuracy on the training data (Chakraborty *et al.*, 2017).

In **constrained optimization** approaches, fairness is formulated as a constraint on parts of the confusion matrix at training time (Mitchell *et al.*, 2020; Caton and Haas, 2020). Both regularization approaches and constrained optimization approaches can be unstable, i.e., small changes in dataset might affect the outcome (Friedler *et al.*, 2019).

Adversarial learning can also be applied with an adversarial model attempting to identify unfairness in the primary model's outputs (Celis and Keswani, 2019). This can also be formulated as a multi-constrained optimization problem (Caton and Haas, 2020). Xu *et al.* (2019) present a more complex adversarial learning approach that attempts to model causal factors with twin generators modeling observed and fair versions of the observed data, with discriminators separating generated from real data and separating the protected and unprotected groups.

For algorithm-in-the loop decision making, Noriega-Campero *et al.* (2019) propose that decision makers can adaptively collect information to ensure fairness for groups and individuals as necessary. This involves an iterative process between modeling and data gathering or preprocessing, so it is not strictly a model approach; it does, however, implement the observation of Chen *et al.* (2018) that different causes or types of unfairness may need different interventions, and some can be addressed by collecting more or less data.

3.3.4 Postprocessing

Post-hoc fairness leaves the data and model alone, but post-processes the model outputs in order to provide fairness. One technique is through **thresholding**, i.e, using different decision boundaries for different groups to ensure non-discriminatory outcome under some definition (Kamiran and Calders, 2012; Kleinberg *et al.*, 2018). We treat the post-processing technique of reranking in more detail later in this monograph, as it is an important tool for fairness-aware information access.

3.4 Wrapping Up

The algorithmic fairness literature has identified a number of different constructs for measuring and reducing unfairness and discrimination in machine learning systems. These constructs are not all conceptually or mathematically compatible; different ones flow from different ethical goals and assumptions about the data, its social context, and the harms to be prevented or ameliorated. For any given application, it is crucial to clearly and precisely describe the problems at play from ethical and/or

legal perspectives, and to select or derive constructs that operationalize a suitable set of objectives. We further cannot assume that any particular fairness objective composes with other parts of the system, or that any particular solution translates cleanly to other problem settings (Selbst *et al.*, 2019). Fairness needs to be defined, assessed, and ensured in a specific problem setting in light of its full sociotechnical context (Dwork and Ilvento, 2018).

We have only been able to provide a very brief introduction to algorithmic fairness in this section. We refer readers to Barocas and Selbst (2016), Selbst *et al.* (2019), Mitchell *et al.* (2020), Dwork and Ilvento (2018), and Suresh and Guttag (2019) for further study.

4

The Problem Space

Information access systems introduce some fundamental twists to problems of fairness and discrimination, making it difficult to directly apply the fairness concepts for other machine learning settings surveyed in Section 3. These difficulties arise from a number of differences, including the addition of multiple classes of stakeholders (Burke, 2017), the rivalrous nature of allocating retrieval opportunities (Introna and Nissenbaum, 2000; Azzopardi and Vinay, 2008; Biega *et al.*, 2018; Diaz *et al.*, 2020), and the immediacy of the interactive feedback loop (Chaney *et al.*, 2018; Khenissi *et al.*, 2020); classification-oriented fairness definitions are not necessarily well-suited to assessing these situations.

In this section we examine, from several different perspectives, the key considerations involved in applying fairness concepts to information access problems. This starts with discussing how the classification fairness constructs described in Section 3 break down when applied to information access. We then discuss unique types of harms accruing in such systems, the potential for fairness concerns across multiple stakeholders, the variety of fairness constructs potentially at work, and the connections between these constructs and the broader space of information access research and fairness, accountability, and transparency

(FAccT) research. We note that not every harm we discuss here has been documented in the wild; we provide citations for as many as possible, but ensuring that information access is equitable requires researchers and developers to proactively engage with possible harms, not only the ones already known. The ACM Code of Ethics (ACM Council, 2018, Section 1.2) states that "avoiding harm begins with careful consideration of potential impacts on all those affected by decisions", and it is our goal in this section to provide a framework to guide that consideration for fairness-related harms in information access.

Fairness is also not a cleanly-defined set of problems with hard boundaries, but rather a lens that encompasses a range of concerns or (potential) harms, as we noted in Section 3. Several problems that information access researchers have long considered can be viewed as fairness problems; for example, work on popularity bias in recommender systems (Celma and Cano, 2008; Zhao *et al.*, 2013; Cañamares and Castells, 2018) has the effect of trying to ensure that the system is fair to less-popular items. There are also many problems, such as incomplete data, that are general problems for information access but take on a fairness dimension when data is missing in a way that disproportionately affects particular people or groups. We take an expansive view of potential fairness problems in information access, with the aim of promoting a wide range of research and development that identifies and addresses inequity and injustice in information systems and their contexts.

The taxonomy we present in this section is not hierarchical or orthogonal, but is rather a set of overlapping lenses or facets through which fair information access may be viewed. The linear nature of writing forces us to impose a hierarchical structure on our treatment of the literature, but any classification of such a contestable problem space is necessarily imprecise and imperfect. We believe, however, that this set of facets is a useful mechanism for understanding the existing literature and for positioning new developments in the broader problem space. Table 4.1 summarizes the dimensions we consider, which we will treat in the following sections, after first contrasting information access fairness with the traditional classifier fairness discussed in Section 3.

Table 4.1: Summary of dimensions for describing information access harms with pointers to relevant sections where applicable.

Category of harm	
Representational harm	4.4.1, 6.1
Distributional harm	5.4, 6.2
Stakeholder group	4.3
Consumers	4.3.1, 5
Providers	4.3.2, 6
Subjects	4.3.3, 6.4
Platform or system	
Multiple groups	4.3.5
Specific types of bias harms	
Direct misrepresentation	4.4.1
Unrepresentative list composition	4.4.2, 6.1
Unfair utility	4.4.3, 5.2, 6.2
Time scale	
Point-in-time	*most sections*
Evolving over time	4.6, 7
Sources of bias	4.7
Imbalanced user data	5.2
Bias in user activity	
Bias in user modeling	
Bias in content production	
Bias in item modeling	
Bias in retrieval and ranking models	
Bias in evaluation methods	
Intervention points	
Data pre-processing	
Adjusting models	5.4, 6.1.3, 6.2.3
Re-ranking	5.4, 6.1.3, 6.2.3
Software process	5.4

Table 4.2: Comparison between information access fairness and the setting typically assumed in classification-oriented fairness work.

Classification	*Information access*
Independent, separate decisions	Ranked results, item decisions affect other items
Subjects receive one decision	Items subject to repeated decisions over time; users get multiple results
Decisions independent of user	Decisions personalized to user
Target outcome construct independent of user	Target outcome (relevance) subjective to user
Data subjects need fair experience	Multiple stakeholder classes may need fair experience

4.1 What Breaks in Information Access?

Unfortunately, the general algorithmic fairness constructs from Section 3, developed primarily in the context of classification for decision-making, do not apply in a straightforward manner to information access for several reasons. In many cases, this is because information access violates key assumptions of the existing classification fairness methods, particularly the assumption identified by Mitchell *et al.* (2020) that the system makes and evaluates its decisions *separately*, *simultaneously*, and *symmetrically*. In that literature, a classifier is intended to support binary decisions, such as granting or denying a loan; we can try to translate this to information access by saying that the information access system is making decisions about whether and where to display an item in response to an information request. Information access has several significant differences, summarized in Table 4.2.

First, **decisions are not independent**, so they cannot be made or evaluated separately. Information access systems typically rank their results: placing one item at the first position in a result list means no other document can occupy that position in that ranking. In any given news result list, only one article will appear in first position. Even if individual items' utility estimates are independent of each other, they

are ultimately resolved into a single ranking to present in response to a user's information need. This ranking may be multi-dimensional, as in a system that arranges rows or carousels of articles, but there is still an ordering and limited opportunity for an article to be in the highest-attention position.

Further, this ranking (or sometimes a set) can be produced in non-obvious ways that depend on more than just relevance. For example, the principle of *maximal marginal relevance* (MMR; Carbonell and Goldstein, 1998) is used to diversify a result list and make it responsive to a range of possible interpretations of a query. Under MMR, once an item is placed in the first position, a second document that is very similar to the first — and just as useful — may lose the second-place slot in favor of a document that brings more diversity to the crucial early positions of the ranking. That is, the utility of an item to an information need is *dependent*. Continuing the news example, once an article has been selected for the first position and its author given priority for potential readership, an MMR approach to topic diversification would likely pick an article on a different story for the second position; so not only does the first-position allocation exclude other authors from that position, it excludes other authors covering the same story from *subsequent* positions in the ranking as well. We can model result list positions as subtractable or rivalrous goods that are allocated to different documents (and their providers), but this does not immediately resolve the problem; it simply identifies it.

Second, **decisions are repeated over time**, a violation of the simultaneous evaluation requirement. Most existing fairness constructs (with the notable exception of literature on fair bandits or fair reinforcement learning) assume that all decisions are made at a single point in time, and do not attempt to account for the system learning and adapting future possible decisions. Some recent work, such as that of D'Amour *et al.* (2020), addresses the need for dynamic considerations of fairness but this has not yet made its way into widely-used fairness definitions. In addition, the same items may be considered for possible ranking multiple times as users interact with the system. If an item is not given a good position in response to the current information request, that does not preclude it from being given a good position

in the next result set. These changes can come because it is more relevant to the next query, user, and/or context; because the system is engineered to apply some randomness to result rankings even when a query is repeated (Diaz *et al.*, 2020; García-Soriano and Bonchi, 2021); because the system has learned more about its relevance through user interaction (Glowacka, 2019); or some combination of all these. The system may put one news article at the top of a user's home page, and a different article for a different user or even for the same user's next visit. Thus, we can speak of fairness *in expectation* based on a system's properties (Diaz *et al.*, 2020), or *amortized* over multiple information requests and results (Biega *et al.*, 2018). This deviation from the standard classification fairness setting provides significant opportunity to overcome the limitations of non-independent decisions, by providing more opportunity for documents to made visible, but it changes the analysis and design of approaches to ensuring fairness.

Third, **decisions are personalized to users**. Returning to the lending example, a classification tool that estimates a potential borrower's risk of default should not return different risk scores or recommended decisions based on which loan officer is currently using the system. Information access systems, however, are often personalized (Liu *et al.*, 2020). This is especially apparent for recommender systems, where most of the value proposition arises from modeling users' particular tastes and identifying products that match them, but many search engines also personalize to their users, as different users have different information needs (e.g. a programmer and a herpetologist likely have different primary interests when searching for "python") and different preferences for information sources to address a particular need. The global implicit aspects of the information need associated with a user mean that two different users with the same explicit dimensions to their need may well be seeking different items. This applies to many of our examples from Section 1.5; news platforms will tailor news recommendations to the user's interests, professional networking platforms will look for job listings that match a candidate's interests; and music services tailor recommendations and playlists to the user's tastes. In practice, personalization relates to repeated decision-making, as we can return different documents to different users, but has some of its own distinct

implications as well, such as the possibly of unfairness in representations of either user preferences or item characteristics.

Fourth, **outcomes are subjective**. While there is a great deal of subjectivity in framing problems and measuring outcomes, typical fairness work assumes an environment in which true outcomes are, in some sense, knowable, at least at some future time. It is assumed that a loan applicant will or will not repay the loan, and this repayment is independent of the bank employee evaluating their application (although it may, in practice, be affected by lender practices such as payment reminders and late payment policies). In many information access contexts, however, the utility of an item usually has at least some degree of subjectivity to it: different users may disagree on the relevance of a document to a query, different users have different preferences for songs or appraisal of their quality, and so on. The goal of the system in the ideal is not to model some external, objective notion of utility, but to model the relevance of the item to a particular user, in a particular context, with a particular query. The same item may need to receive different scores or ranking decisions in response to different requests either from different users or the same user in different contexts.

Fifth, **multiple stakeholders have fairness concerns**. While most problems have multiple classes of stakeholders — the bank wants to make a net profit, the loan applicant wants a loan, and the community has an interest in residents having access to credit while maintaining low incidence of foreclosure-induced homelessness — the applications considered in much work on algorithmic fairness have a clear stakeholder class for whom fairness is taken to be important. We attempt to ensure that lending is fair to loan applicants, but do not spend effort on ideas of fairness to banks; likewise job applicants and the employer. Information access, however, has multiple stakeholders with salient fairness concerns. Two particular ones, which we introduced in Section 1 and discuss in more detail in Section 4.3, are *consumers* (the users of the system who consume information or products) and *providers* (people or entities who create or provide the items consumed). Research (and system deployment) accounting for fairness in information access needs to identify one or more stakeholder groups for whom it will consider fairness, and to clearly document this decision and its implications. The

exact position and fairness requirements of individuals may also change from product to product, even within the same domain or platform; for example, prospective employees are consumers and employers are providers in a recommender for job listings, but the roles are reversed for a job candidate search tool for use by recruiters. The same platform may well provide both of these types of recommendations.

These differences — and likely others — mean that it is important to study fairness in information access as a distinct problem in its own right. The extensive work to date on fairness in other algorithmic settings has much to say that can and should inform this work, but we cannot expect methods or metrics from classifier fairness to directly and immediately apply to information access without adaptation to the particularities of information access problems.

Machine learning fairness has much to teach us about how to frame, understand, and measure fairness, but the details often do not directly translate to information access problems. Naïve application of classifier fairness constructs to information access often breaks down.

4.2 Kinds of Harms in Information Access

There are a number of different ways an information access system can harm one or more of its stakeholders, particularly arising from its role as a mediator of users' access to information that may or may not meet their needs.

An unfair distribution of performance — specifically one that favors over-represented populations — can systematically hurt the retention of entire subpopulations of users. An information access system that optimizes for mean performance can improve mean performance metrics (e.g. user satisfaction), at the expense of under-performance on under-represented groups, which are dominated by over-represented groups; metrics may also improve through the attrition of users from an under-represented group. In simulation experiments, Hashimoto *et al.* (2018) demonstrate that traditional empirical risk minimization results

precisely in these dynamics. In the context of two-sided recommendation, attrition may occur from content consumers or providers, potentially compounding any effects over time (i.e. attrition of providers can impact the size of the catalog and performance for consumers).

Increasingly, especially in protected domains like housing (HUD, 2020) and employment (Raghavan *et al.*, 2020; Sánchez-Monedero *et al.*, 2020), legal human rights regulation is being expanded to include algorithmic decision-making, like information access (Wyden, 2019; Thune, 2019). As such, techniques for auditing and addressing systems for unfairness will become important from a legal perspective.

Journalistic investigation provides an alternative way in which algorithmic unfairness may be surfaced (Diakopoulos, 2015; Angwin *et al.*, 2016; Pelly, 2018). These investigations can provoke regulation and hurt user perception and trust, potentially leading to attrition.

Besides the utilitarian effects on the information access provider, harms include a variety of social externalities. Metrics such as user satisfaction or retention can ignore the broader impact of mediating information, especially news, which can affect social and political institutions. While content consumers can be affected by mediation, the livelihood of content producers can particularly be at the whim of algorithmic decision-making, incentivizing classes of content more amenable to distribution than under-represented content.

These harms suggest that information access system designers should understand the broader implications of the technology they produce. Indeed, early in the development of information retrieval, Belkin and Robertson (1976) noted the ethical responsibility of information retrieval researchers to avoid political or economic manipulation. Librarianship provides a professional discipline close to those designing information access systems and a code of ethics that ensures "equitable services are provided for everyone whatever their age, citizenship, political belief, physical or mental ability, gender identity, heritage, education, income, immigration and asylum-seeking status, marital status, origin, race, religion or sexual orientation" (IFLA Governing Board, 2012).

There are multiple ethical, legal, and business reasons why information access system developers should consider fairness in their system design and evaluation.

4.3 Fair for Who?

While most of the history of information access research has concentrated on optimizing system performance for user outcomes (see, for example, the many rounds of TREC providing evaluations of IR systems' ability to retrieve relevant results, and the standard practice of assessing predictive or top-N accuracy in recommender systems), there has been a growing acceptance in recent years that, in some contexts, information access systems serve multiple goals and possibly multiple parties, each of which is affected by the system's results and behavior. The integration of the perspectives of multiple parties into recommendation generation and evaluation is the goal underlying the sub-field of *multistakeholder recommendation* (Abdollahpouri *et al.*, 2020; Mehrotra *et al.*, 2018).

Many e-commerce sites operate as **multisided platforms**, a business model analyzed in the economics literature by Rochet and Tirole (2003) and Evans *et al.* (2011). One important finding is that different applications require different distributions of utility. In many multisided platforms, there is a 'subsidy side' of the transaction where one set of parties uses the platform at a reduced cost or no cost. For example, users of the OpenTable restaurant reservation service do not directly pay for reservations; instead, restaurants pay for each reservation made (Evans and Schmalensee, 2016).

In information access as well, the outcomes of the system may be biased towards one group of parties for similar reasons. In addition, the need for personalization may vary across systems and between stakeholders. For example, in an e-commerce site, product suppliers will usually not care about the characteristics of consumers: as long as products are surfaced to likely buyers — either through recommendations or search results — they will be satisfied with the behavior of the system. Online advertising is different: typically an ad campaign is targeted towards a

particular audience, so a recommended ad placement is only considered successful if the ad matches the user's interests, to the extent they are known, and the user matches the target audience towards which the campaign is oriented.

Any system may have a multiplicity of stakeholders who are impacted by its decisions. In the case of fairness, the most salient ones will often be those noted above: consumers and providers. It may well be the case that these stakeholders individually have no interest in fairness; they may be primarily interested in the best outcomes for themselves. Fairness is usually a constraint that the system creators impose on outcomes, either to satisfy their own organizational mission or to meet the demands of yet other less-proximal stakeholders, such as government regulators or interest groups. A few counterexamples do exist in the form of technologies to allow stakeholders to prevent certain kinds of unfairness even if the platform owner is accepting of unfair outcomes (Nasr and Tschantz, 2020; Kulynych *et al.*, 2020).

4.3.1 Consumers

Fairness towards consumer stakeholders may be grounded in different normative concerns: a goal of beneficence and the avoidance of various types of harms; a basic sense that the metric of system performance should include the broad distribution of its benefits; or, a more practical concern that unsatisfied users may go elsewhere.

We can characterize consumer fairness in various ways. The most straightforward is through quality of service, particularly in terms of error, user satisfaction proxies, or other performance metrics (Section 5.2; Mehrotra *et al.*, 2017; Ekstrand *et al.*, 2018b). If particular classes of users less utility than others, whether measured in terms of prediction accuracy, ranking accuracy, or other measures, then we may say that the system is treating those users unfairly.

A system may also be unfair if its output is discriminatory in the content provided to different groups. For example, it is well-documented that real estate agents in the US have regularly steered minority home buyers to a limited set of neighborhoods; such incidents prompted the US Fair Housing Act, which explicitly disallows such activities (Rothstein,

2017). Not to run afoul of such laws, a real estate recommender would need to be sure that its lists of recommended properties contained a fair distribution of opportunities regardless of the buyer's minority status. Thus, fairness may involve the specific content provided rather than differential performance on some error metric. This issue can arise in many other domains as well, such as ensuring job seekers of different genders or ethnicities have access to comparable job listings; recommending lower-paying jobs to some groups of users than others would violate this principle, and in some jurisdictions may be illegal (case law is not yet settled on this point). This phenomenon has been studied, for example, in Facebook's ad targeting platform (Ali *et al.*, 2019).

Finally, groups of consumers may be impacted if they have to incur disproportionate costs in order to use a system. Such costs might come in the form of information disclosure or effort. For example, a user with a disability may find they have to set up specific filtering rules to get the recommender system to provide acceptable hotel room recommendations and an able-bodied user does not, in spite of having provided the system with a similar amount of preferences or ratings. Systems may perform better if users opt to share more data, or under-perform for users who are on low-quality connections and cannot provide as much data about their information needs (will note here that it is not necessarily possible to fix every potential fairness problem!). It is also not clear that privacy-preserving recommender systems impose equal accuracy or quality costs on different groups of users (Ekstrand *et al.*, 2018a; Bagdasaryan *et al.*, 2019). Users who experience marginalization may also experience the tracking of activity and the generation of profiles that comes with personalized recommendation as considerably more threatening than others (Browne, 2015; Burke and Burke, 2019).

These concepts of fair treatment do not necessarily correlate. A system may look fair in that different groups of users receive comparable quality of service as measured by system effectiveness metrics, but perform poorly at presenting protected group users with the "best" inventory. One specific way this can manifest is when using clicks to measure information satisfaction: users click on results presented to them, so the system may appear to be delivering satisfactory results,

when in fact one user would prefer to receive the results another user is receiving. Thus, a system designer looking to protect consumer fairness will need to think carefully about the types of harms user might experience and how to detect and measure them.

4.3.2 Providers

There is a fundamental asymmetry between the various stakeholders participating in an information access system. Consumers come to the system to find information and are thus active in the information- or product-seeking process ('lean back' recommendation experiences aside). They may be able to get multiple sets of results if they wish. Content providers, on the other hand, have a more passive role: their items are presented when and if appropriate users arrive, and they typically have little control over the recommendation or retrieval function. Despite this asymmetry, **fairness** concerns can arise in similar ways. Different groups may experience greater error when predictions of their items are made. For example, the system may systematically under-estimate user preference for books by minority authors (Yao and Huang, 2017).

More often, however, the concern will be about the exposure of items in results. This notion of provider fairness is concerned with how different providers, either individually or as members of protected groups, have their items appear (or not) in the rankings produced by a particular system. For example, in a search or recommendation tool to help recruiters find candidates for an opening, we want to ensure that the candidate lists it produces treat protected groups fairly (Geyik *et al.*, 2019). The personalized nature of recommendation (and many search systems), the fact that individual result lists are limited in size, and the rank-ordered nature of most information access systems means that we can only hope to achieve this kind of fairness over time and across multiple user visits (Biega *et al.*, 2018; Diaz *et al.*, 2020). In some cases, a provider's items might be a poor fit to the users of a particular system. Consider a pianist on a job site whose primary user base is carpenters and electricians: there might not be many recruiters for whom such an applicant is relevant, and presenting their profile is likely unhelpful. All of these considerations complicate the problem of measuring and ensuring provider fairness.

As we have used it so far, the term "provider" is a simplification of what can be quite complex systems of production and distribution that produce the documents or items in a system's inventory. The area of popular music is a good example, where the beneficiaries of the recommendation of a music track can be quite diverse, from the artist whose name is on the track, to the other musicians involved in the recording, the songwriter(s), the producer, the record label, etc. Fairness may not mean quite the same thing to each of these individuals; unfortunately, little data is currently available on how different provider groups perceive the fairness of recommender and information retrieval systems specifically. In one recent result, Ferraro *et al.* (2021) report on interviews with musicians and their perceptions of fairness in music streaming platforms, with special attention on female artists. However, most literature focuses on the perceptions of recommendation consumers, and filling the gap to understand providers' experiences is an area in need of significant study.

There is significant conceptual overlap between provider-fairness and the diversity of search and recommendation results (Ziegler *et al.*, 2005; Steck, 2018). However, when diversity is invoked as a desirable property of an information access system, it is usually in the service of some user-oriented goal. For example, in information retrieval, query aspect diversity can compensate for the fact that the system may have an incomplete understanding of an ambiguous query, and covering multiple possible aspects increases the chances that one of them is correct. Fairness as a social justice concern seeks varied outputs for completely different reasons.

Provider fairness in this sense also has a strong connection to ideas of fair allocation from welfare economics (Moulin, 2004; Thomson, 2016). The resource at issue is the opportunity for a provider's item to be made visible to a user, and the question is how to allocate that resource among various providers and what are appropriate desiderata. This is a cornerstone topic in social choice and has found practical application in a number of areas including allocating courses in schools (Budish and Cantillon, 2012), papers to reviewers (Lian *et al.*, 2018), and numerous other settings (Roth, 2015; Aziz, 2019).

4.3.3 Subjects

Retrieved items can sometimes themselves be about individuals, which we call **information subjects**. A well-known example arises in image search and recommendation: in many systems, image searches for terms like "CEO" turn up results that over-represent white men (Metaxa *et al.*, 2021). Female and non-white information subjects are not directly materially disadvantaged, but the results give a false impression that leads to the perpetuation of stereotypes (Noble, 2018) and possibly associated loss of opportunity. Karako and Manggala, 2018 examine this phenomenon and provide methods to address it, with a running example of seeking to provide a set of workout images that broadly represent the population. Similarly, news recommendation may fail to give balanced coverage of issues affecting different groups, such as rural vs urban residents.

Another example of possible subject unfairness arises in medical information access. If the studies returned in response to a query for current research on a medical condition a doctor is treating do not report on experiments whose subjects are not representative of the population — or particularly do not include people sharing the patient's medically-relevant characteristics — information may be missing for providing the best outcomes for the patient. We are not aware of significant research on this particular potential problem, but proactive study of information access equity requires that we consider it.

Technically, subject fairness (being fair to information subjects) has a lot in common with both provider fairness and diversity. In each case, it is the items over which representation is sought. However, in seeking subject fairness, it may be even more important that individual lists be diverse, as the goal of diverse representation is not necessarily satisfied by alternating between diverse and non-diverse lists, something that might be acceptable in a provider fairness setting.

4.3.4 Side Stakeholders

The concept of multistakeholder recommendation (Abdollahpouri *et al.*, 2020) includes many stakeholders beyond those we have discussed, many of whom may have fairness concerns that a system should address.

Indeed, in some settings, regulatory agencies may be the most important stakeholders for deciding the minimum legal standards with respect to fairness and / or non-discrimination that a system must meet. In other cases, the structure of a platform entails the participation of stakeholders who are neither consumers nor providers, but are still impacted by specific parameters of transactions on the platform.

For example, consider a food delivery platform such as UberEats[1] discussed by Abdollahpouri (2020). This platform uses recommendations to match consumers with restaurants where they might order food to be delivered. The deliveries are made by Uber's drivers. There may be fairness concerns relative to the consumers or providers here, but there may also be fairness concerns over the set of drivers. These individuals do not participate in the recommendation interaction but the recommendations may impact them. For example, a goal might be to ensure that protected groups among the driver population do not receive fewer orders than others or do not receive a disproportionate number of difficult and/or low-tip jobs.

Another example arises in vehicle routing, which can be viewed as a kind of information access system where the items are routes or route segments. The routing systems built in to mapping platforms such as Google Maps and Waze allow users to build routes for a variety of objectives, including "beauty" and avoiding heavily-traveled routes. These systems can have the effect, however, of increasing traffic on side streets and through residential neighborhoods; residents have a significant stake in these kinds of changes to traffic patterns (Johnson *et al.*, 2017; Fisher, 2022).

There has been comparatively little published work to date that considers such "side stakeholders" who are indirectly impacted by information access, but it is an important direction for future research.

4.3.5 Joint Fairness

So far, we have only considered each group of stakeholders in isolation. This is the starting point for much machine learning fairness research, but it is a simplification. In practical settings, multiple groups may

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

experience harms and benefits through the actions of an information access system and therefore multiple simultaneous concerns may arise (Mehrotra *et al.*, 2018).

Information access systems are multisided platforms as noted above and therefore it is possible that the consumers of results and those provider side may each have fairness concerns. For example, consider a recommender system for rental apartment listings. Fairness concerns with respect to renters are well-established in housing anti-discrimination law; a system should not discriminate in the types of listings it provides on the basis of protected attributes like ethnicity or religion. But at the same time, there could be concerns relative to landlords. Hypothetically, if a system were found to be steering potential tenants with poor credit histories to properties owned by minority landlords and better prospects to other landlords, this would also be discriminatory.

Multiple fairness concerns can also arise on a single side of the information access interaction when there are multiple groups to consider. For example, a search engine may misrepresent both women and ethnic minorities in the results from image searches. The idea of fairness among multiple protected groups has seen some initial exploration under the label "subgroup fairness" (Kearns *et al.*, 2017; Kearns *et al.*, 2019; Foulds *et al.*, 2020), but there is still much more to do. In particular, existing work as yet does not take into account the particular, compounded, challenges that may be encountered by individuals at the intersection of multiple protected categories (Cho *et al.*, 2013).

4.3.6 Cross-Group Harms

In addition to unfairness for particular stakeholder groups directly harming that group's members, unfairness for one group may also harm other groups.

If a system is provider- or subject-unfair, it may provide consumers with skewed perceptions of the space of content providers or subjects (Noble, 2018). A system that under-exposes job candidates from racial minorities may lead its users to believe such candidates are less common than they actually are. It may also make it difficult for consumers to find content they particularly like, because such content is created by providers who are not well-represented by the system.

If a system is consumer-unfair, it may under-serve — and thereby discourage — a provider's primary audience, making it difficult for that provider's content to find an audience.

Measuring and countering unfairness in an information access system requires clearly identifying **who** is being considered in a particular evaluation or intervention. Different groups have different concerns that will give rise to different metrics and techniques.

4.4 Fair How?

Another way of understanding fair information access is to look at *how* different participants in the system may experience or be harmed by unfairness. As with general algorithmic fairness, information access stakeholders may experience unfairness on individual or group bases.

Classical welfare economics examines fairness in the form of distribution (Moulin, 2004): how to divide a resource fairly among individuals, all of whom have some claim to it. This type of fairness consideration can be considered distributional. In the information access context, there are different multiple resources in question, depending on the stakeholder: providers and subjects receive exposure with its resulting material and reputational benefits, and consumers receive information that hopefully meets their information needs.

Crawford's representational harms also affect information access systems, when the system misrepresents a user, an item, or the information space. We distinguish between a representational harm, where the harm itself is one of misrepresentation (e.g. misgendering a book author, or presenting results that discourage girls from seeing themselves as possible CEOs), and unfairness in the system's internal representations (e.g. user or product embedding spaces having a stereotyped component). The latter may result in any of the kinds of harms we discuss in this section (either representational or distributional), or may be compensated for by other aspects of the system's behavior.

Sometimes harms, particularly with respect to legally-defined protected attributes, will be defined and proscribed by law; others will be a matter of policy for the designer of a particular system. This breakdown is also not entirely crisp: some harms will fall under multiple categories simultaneously. We submit that it nonetheless provides a useful way for understanding the ways in which discrimination and related problems manifest and harm the system's participants.

4.4.1 Direct Misrepresentation

An information access system can cause direct representational harms when it presents inaccurate information about items.

Direct misrepresentation of item or provider characteristics can harm both consumers and providers. Consumers are harmed because they obtain inaccurate information, and the misrepresentation may keep them from finding content through systems such as faceted browsing interfaces or detailed keyword searches. Providers are harmed first because they are misrepresented, which can be harm in itself, but also may not have their product accurately discovered. For example, if a children's book is not correctly labeled as such, then users may not find it when they are browsing or searching for children's books on a topic. This specific harm is also an example of a multi-category harm, as it is also an unfair allocation of exposure to the book and its author and publisher.

Direct misrepresentation can itself be unfair (as in the example of misgendering) or it can happen in an unfair way (for example if some providers' content is more likely to be correctly represented than others in a systematic way). It can also be either individual (if similar items do not have similar representation) or group (if socially-salient groups of items or producers are systematically misrepresented). For example, as of 2020, a system using book author data from the Virtual Internet Authority File (VIAF) will exhibit group-based direct misrepresentation of transgender and non-binary authors, because multiple and non-binary gender identities are not accurately stored in the VIAF (Ekstrand and Kluver, 2021).

Misrepresentation can also harm broader sets of stakeholders. For example, Nagel (2021) documents an instance of a search engine result page showing a carousel labeled "Famous Cherokee Indians" and displaying photos of several celebrities, most of whom have no documented Cherokee ancestry or affiliation. This misrepresents the celebrities themselves, but likely does not cause them significant direct harm; its more substantial impact is misrepresenting what it means to be Cherokee — and therefore the Cherokee nation — to users of the search engine. This is also an example of a stakeholder that is impacted by an information access system but is not a producer, consumer, or subject of its items. While the root cause of this problem is likely missing information in the underlying knowledge graph, it has an effect that compounds the difficulties already faced by indigenous communities; fairness-related problems can stem from any of the many issues in data or algorithms that affect other aspects of information access systems.

One additional potential harm that can arise from representation is that activating stereotyped perceptions a user may hold can affect their processing and assessment of information (Bodenhausen and Lichtenstein, 1987); inaccurately representing producers or subjects, or even representing them accurately but unnecessarily, in a way that connects with users' negative stereotypes may impeded their ability to accurately and appropriately make use of the results the system provides, particularly in complex assessment situations.

4.4.2 Unfair Result Set Composition

An information access system can exhibit unfairness in the composition of its result sets and rankings. This can have further downstream effects, as information access systems are often the first step in users' quests to gather information for other purposes.

One example is the previously-mentioned case of image search results for "CEO" (Crawford, 2017). The set of results can affect users' perceptions of both the *current state* and the *possibilities* of the role of CEO (and it isn't immediately clear which one to emphasize if they differ — should the gender representation in such a result set reflect the set of fortune 500 CEOs, all CEOs, or the general population?). Further,

if a student searching for "CEO" images for preparing a presentation, this not only affects who they see in the role of CEO, but affects the images available to them for communicating with their peers. This kind of downstream effect can arise in many settings, in education, business, and beyond. Any unfair representation that results in reinforcing stereotypes, either through the selection of items or explanations of results, may amplify — or at least perpetuate — societal biases. Hoffmann (2019) argues that this kind of representation of what is "normal" has significant impact on how we understand and navigate the world and on the dignity of the people represented: in the situations we discuss in this monograph, the creators and subjects of information resources. Such unfairness is also not limited to representation of the people involved; Raj *et al.* (2021) notes that system results can also perpetuate gender stereotypes, which is of particular concern when used by children.

Unfair result sets can also arise through a personalized information access system's user profile, or its interactions with item representations. A movie recommender may emphasize item relationships along gender-stereotyped lines, so that a user receives "guy" movie recommendations based on a few movies they've watched, instead of a set of recommendations more broadly reflective of their tastes.

Ways result sets can have unfair composition include (Noble, 2018):

- Reinforcing stereotypes (of users, content, providers, subjects, or any combination)
- Presenting an inaccurate picture of the information space
- Biasing users' sense of the possibilities of the information space

4.4.3 Unfair Distribution of Benefits

Perhaps the most obvious way in which an information access system can be unfair is by being unfair or discriminatory in how it distributes benefits to its stakeholders. Information access provides consumers with information that hopefully meets their needs, and providers with opportunities for their content to be discovered; this can have many repercussions, including financial (if access directly or indirectly results

in revenue) and reputational (by their content being made broadly known), and can have material impact on providers' career prospects.

Distribution can be unfair at an individual level, if similar users do not receive similar quality of results, or if similar content does not have similar opportunity to be presented to users. While similarity in the general case is often difficult to assess, in information access we often have some estimate (or even measurement) of content's relevance to an information need. Relevance assessments provide a useful basis for similarity-based evaluation of the distribution of opportunities for user attention to content providers is individually fair. If two content providers create documents that are comparably relevant to a particular query, and they do not receive comparable exposure in result lists or engagement from users, we may say that the system violates such an individual fairness objective (Biega *et al.*, 2018).

Utility estimates themselves may be unfair, if individual items with comparable characteristics with respect to their ability to meet the user's information need do not receive the same score. Such cases may require additional attention to ensure individual fairness. Individually-fair distribution of result quality is, to some extent, already addressed by information access evaluations that consider the distribution or order statistics of quality and accuracy metrics, but differences in the rate at which attention and relevance drop off mean that static rankings generated by the Probability Ranking Principle are not necessarily fair, whereas stochastic ranking policies that attend to the distribution of exposure or attention can correct this discrepancy (Diaz *et al.*, 2020).

Group-based distributional discrimination may arise in different forms and for different stakeholders. On the consumer side, systematically underserving groups of users and failing to capture their perspectives or interests is a form of distributional group unfairness (Mehrotra *et al.*, 2017; Ekstrand *et al.*, 2018b); we discuss this in Section 5.2. It can also show up in less obvious ways, such as failing to properly interpret queries from children (Dragovic *et al.*, 2016).

Provider-side group-based distributional unfairness has the clear manifestation of under-presenting results from particular, often disadvantaged, groups of content providers, such as authors who are members of ethnic or gender minorities. Section 6 will cover this space in much more detail.

In addition to the kinds of protected and sensitive groups typically considered in fairness work, such as gender, religion, race, and ethnicity, information access applications bring additional sets of groups towards which we may want to ensure distributional fairness. The work on *cold start* in recommender systems (Schein *et al.*, 2002) can be framed as ensuring that new users and items are fairly treated. We may want to ensure new or independent authors, artists, or studios aren't crowded out by more established sources, reducing the ability of up and coming providers to thrive. Ensuring good quality across different languages or regions can also be beneficial in information access applications.

There are important distinctions between distributional fairness of different resources for different stakeholders that affect how we measure and provide fair distributions. One key distinction is the *subtractibality* (or rivalrousness) of the resource in question (Becker and Ostrom, 1995): does one person's use of the resource affect the ability of others to enjoy it? For provider-side fairness, positions in result rankings are clearly subtractible: in any given ranking, only one item can be placed in the first position, and placing it there denies the position to other items. The system may provide fairness overall by placing different items in the first position in different rankings, possibly even in response to the same information need, but any individual opportunity for user exposure (defined as a particular ranking position in an information access transaction) is a subtracible good. Chakraborty *et al.* (2017) lean on this formulation by adapting existing algorithms for sharing limited resources — specifically CPU time — to help fairly allocate recommendation opportunities in sharing platforms.

On the consumer side, subtractibility is less clear. System utility is typically not subtractible: one person obtaining high-quality, useful responses to their information need typically does not generally prevent others from receiving similarly relevant results. Many classes of items are also non-subtractible, including digital content (web pages, streaming music, etc.) and plentiful physical goods where user demand is not likely to exhaust supply. Other items, however, are subtractible: online auctions, for example, often have very limited stock, and only a small number of applicants will actually receive any particular job;

while simply presenting the item to one user in a result list does not itself reduce the ability of others to benefit from it, successfully consuming the item does. This is also a significant concern for reciprocal recommendation contexts such as those in matchmaking platforms for dating, mentorship, and other involved relational commitments (Pizzato *et al.*, 2010), as "people have limited availability, so one person should not be recommended to too many others". Patro *et al.* (2020) apply the concept to local business recommendations, which often have limited physical space that is further reduced by distancing requirements for public health; if too many people are recommended the same restaurant, they may not all be able to enjoy it safely, but spreading out the recommendations can help more people get to a restaurant that they can enjoy. Whether or not it is necessary to reduce the resource distributed to some stakeholders in order to improve it for others is a key concern in describing the distributive harm that an information access system should avoid.

As much of the existing literature in fair information access focuses on distributional aspects of unfairness, we go into this point in much more detail in later sections of this work (in particular, consumer distributional fairness in Section 5.2 and provider fairness in Section 6).

Fairness in information access also requires careful attention to **how** the stakeholders may be harmed. Again, different (potential) problems require different approaches.

4.5 Fair from What Vantage Point?

Orthogonally to the stakeholder group, we can consider different vantage points for measuring fairness. As noted above, there is a range of different metrics and methodologies for measuring information access system performance and these give rise to different ways of thinking about fair outcomes.

Since information systems often have an underlying predictive element, we can measure the accuracy of its predictions: does the system

predict what items a user will like and how much? Accurate results deliver utility to users (helping them find items of interest) and also to item providers, because it means that their items are reaching appropriate targets. As an element of fairness, we might ask then if the system delivers different degrees of accuracy to different groups of users, serving some well and others poorly, and/or if the system differs in its prediction accuracy across protected groups.

There may be other aspects of information access performance of interest, depending on the recommendation application. In some settings, it may be possible to rank items on some objective scale of desirability. For example, credit card offers with lower interest rates and lower fees, and offer larger credit limits are better than those that charge higher interest and fees for less credit; all other things equal, a job with a higher salary is better than a lower-salary job. The objective quality of the contents of a delivered results list is therefore an element of utility, especially for consumers. We may want to measure the comparative quality of such lists across protected and unprotected groups to identify possible discrimination. On the provider side as well, some products may be more profitable than others and unfairness may take the form of presenting low profit items for one group and high profit items for another.

In other cases, we may not distinguish between the qualities of items but rather their relative frequency of appearance on lists. A provider whose items appear very infrequently on result lists may feel that the system is being unfair in not promoting their products. This may also take the form of unfair representation as discussed above. From a consumer's point of view, differential distribution of item appearances may also have an element of unfairness: consider, for example, a recommender system that presents science toys on lists shown to boys but not on those shown to girls (Raj *et al.*, 2021).

Fairness concerns may go beyond the distribution of quantitative exposure or utility.

4.6 Fair on What Time Scale?

Typical off-line evaluation methodologies for information access systems treat the test set over which outcomes are measured as if all the results are generated at the same time; this corresponds to the *simultaneous* assumption articulated by Mitchell *et al.* (2020). This tests how the system state induced by the training data produces results for all stakeholders at a point in time, but does not necessarily reflect how users, producers, and other stakeholders actually experience the system's effects. Lathia *et al.* (2009) and others have looked at evaluation over time, and Sun *et al.* (2020) advocate using time to split training and test data to provide a more accurate picture of system performance, but we have not yet seen significant applications of these concepts to studying fairness.

Since information access systems in practice deliver results lists to users who arrive at the system over time, it is important to examine fairness as a property of information access outcomes delivered over time (Biega *et al.*, 2018). For example, we might ask whether a particular fairness metric has been achieved within a set of results delivered over some time window Δt. A system might look back over such an interval, determine whether a fairness metric has been met or not, and try to adapt its algorithm to deliver improvements over the next interval (Sonboli *et al.*, 2020a).

Finally, the question of the dynamic nature of information access delivery leads to the question of the impact of results on user behavior, the impact of that behavior on subsequent system learning (the feedback loop). As a simple example, we can consider the impact of popularity bias, the fact that many recommendation algorithms reinforce the popularity distribution across items (Jannach *et al.*, 2015). A popular item is recommended, then experienced and rated, appearing more popular in the data, and getting recommended more often, etc. While the unequal distribution of popularity is natural, this kind of positive feedback loop can exacerbate such distributions and associated unfairness (Chaney *et al.*, 2018).

The repeated nature of information access and its evolution over time, particularly as the system learns and updates its models in response to user interactions, means that point-in-time analysis is not sufficient to fully understand the fairness-related behavior of the system.

4.7 Fairness and the System Pipeline

In addition to the various dimensions along which we can define what it means for an information access system to be unfair, this unfairness can arise from a variety of sources. Section 2.1 and Figure 2.1 described several components of a typical information access system; any of these components introduce unfairness to the system, and we can consider both evaluations and fairness interventions at many stages. In most stages, unfairness can arise from the underlying data involved in that stage, the computational models used to make access-relevant inferences from that data, or both.

Item understanding can affect both producer fairness and the ability to correctly locate documents to meet an information need. Unfair representation of documents or categories may introduce representational harm or have downstream distributional effects. In a job search scenario, the document collection used to build the information system may have predominantly male candidates; if there are gendered aspects to how candidates present themselves in their documents, the system may learn gender correlations to job capabilities. Unfairness may arise from calculated metadata as well: for example, unfair inference of sentiment scores from review documents may introduce unfairness to the system at the classification stage. Although it is impossible to have a holistic understanding of the items in a retrieval system, it is necessary to be intentional in mitigating potential unfairness and highlight possible limitations.

Representation learning, and possible unfairness or discrimination in learned representations such as content embeddings, is one specific way in which item understanding may contribute to unfair outcomes

from an information access system. For example, the presence of racial bias in reviews (Speer, 2017) or gender bias in a job description can lead to an unfair outcome to different groups of stakeholders due to the bias in representation. Bender *et al.* (2021) provide an extensive discussion of the problems that can arise specifically when using language models for item understanding, a number of which touch on fairness.

User understanding most directly affects consumer fairness. The system needs to understand who its users are, their information needs, and their capabilities and preferences; the user modeling dimensions of information access particularly focus on this, as does the query understanding component of information retrieval. The system does not necessarily learn these characteristics in an unbiased way.

Retrieval and **rendering**, often including ranking, are central to the observable output of the information access system. In Section 6 we will go into more detail on the problem of *fair ranking*, which often connects to re-ranking ideas from information retrieval. Other aspects of rendering, such as result presentation, are less explored from a fairness perspective. Several biases may get introduced in the system purely from the user experience of the result page. Much more research is needed to understand how different design elements may introduce unfairness to otherwise fair retrieval results.

Behavior understanding is how the system improves itself, either through automatic learning or feedback to system designers. However, as noted previously, that opens up the possibility of creating a feedback loop of that reinforces and amplifies any unfairness in the system. The varied nature of behavioral feedback between user groups may also affect the system's ability to accurately learn to provide relevant results towards different groups. Relevance feedback and click-through data are common ways of understanding user behavior and improving information access systems, but the data underlying them can easily be biased with respect to any of the stakeholders. Such feedback loops can easily amplify small differences, such as one item being slightly more relevant than another, into large differences in exposure or attention (Ensign *et al.*, 2018).

Evaluation of information access systems, as discussed in Section 2.5, is inherently different from the evaluation of classification systems. Where classification has a rather fixed notion of decisions and outcomes,

evaluating retrieval systems relies on understanding the underlying user model (Singh and Joachims, 2018; Biega *et al.*, 2018; Sapiezynski *et al.*, 2019). Evaluation based on biased relevance judgements or user response data can result in incorrect design decisions with fairness implications, or in selection of models or parameters that are unfair; evaluation is also a crucial place for assessing the fairness of an information access system. One approach to this latter goal to evaluate fairness and relevance separately (Yang and Stoyanovich, 2017; Das and Lease, 2019). There is a scope for novel metrics that incorporate different notions of fairness as well as relevance.

Unfairness can arise from both data and models at any stage of the information access process. Much research is needed to understand the role each plays in the overall fairness-related behavior and impacts of an information access system.

4.8 Fairness and Other Concerns

The relationship between fairness and other concerns and concepts in information access is not straightforward. As we have argued, some historical concerns can be framed as certain kinds of fairness problems: recommender system cold-start work, for example, seeks to ensure that new items and users are given a fair opportunity for exposure or quality from the recommender system, and long-tail recommendation looks to prevent the system from being unfairly biased towards popular items disproportionate to their utility to users. Length normalization attempts to decrease systems' unfair preferential treatment of long documents (Singhal *et al.*, 2017). While these do not deal with the kinds of socially-salient groups often considered in fairness research, they still represent the core logic of anti-discrimination: items should be retrieved based on their utility to the user's information need, not other incidental factors such as popularity, newness, or length (except when those factors are directly related to the need). Beyond the desire to root out unwanted incidental biases, fairness research can be viewed as extending this logic from biases stemming from endogenous properties of items and

their representation in the system to biases stemming from exogenous properties that relate to the broader social location of items, providers, users, and other entities.

Other concerns are related, and may have metrics and technical machinery that can be reused for fairness purposes, but flow from different normative concerns. One frequently-mentioned example of this is diversity: provider fairness and diversity look very similar, and systems providing more diverse search results or recommendations will probably often be more fair towards different providers. However, they flow from different normative concerns and should therefore be assessed with metrics that reflect those concerns. Diversifying techniques such as MMR (Carbonell and Goldstein, 1998) or xQuAD (Santos *et al.*, 2010) can be used to improve fairness as done by Sonboli *et al.* (2020b), but the results should be assessed on the basis of fairness.

Finally, some concerns may be in tension. Work in both information access fairness and general ML fairness often discusses a tradeoff between fairness and accuracy. Sonboli (2022) makes the terms of this potential tradeoff precise by framing it specifically as a tradeoff between fairness and accuracy in classical metrics used to measure recommendation accuracy in offline evaluations. Wu *et al.* (2021) treat this tradeoff itself as a fairness problem, arguing that decreasing consumer utility to improve fair provider exposure (Section 6.2) is unfair to consumers, but prioritizing consumer metrics with no consideration to equity of exposure is fair to providers (and they provide definitions and algorithms for navigating this multi-sided fairness).

However, the well-known disconnect between these offline metrics and online metrics of utility or user satisfaction (see e.g. Rossetti *et al.*, 2016; Kouki *et al.*, 2020) means that the relationship or tradeoff between fairness and utility may be very different than the relationship between fairness and offline accuracy. Even in the offline setting, though, the tradeoff is not necessarily inherent, as Bigdeli *et al.* (2021) show empirically through the existence of techniques improving both accuracy and fairness; Dutta *et al.* (2020) argue that observed tradeoffs often arise due to bias in the data used to evaluate accuracy. There is not currently enough research on the contours and limits of either fairness, accuracy, or utility to make definitive conclusions on their relationship in the general case.

Fairness may be in tension with accuracy or utility in some cases or experimental settings, but more research is needed to more fully understand and predict their relationship. Fairness has significant overlap or complementarity to other concerns for information access, such as diversity and popularity bias.

4.9 Contributing Back to ML Fairness

In light of the significant differences between information access fairness and the fairness contexts typically studied to date, and the various challenges in the problem of ensuring fairness in information access systems, we also believe that information access has much to contribute back to the broader algorithmic fairness community.

One contribution is that the ways information access violates assumptions of classical fairness techniques (Section 4.1) can help make the limits of those techniques concrete. We can point to specific applications where decisions can no longer be made and evaluated separately, or simultaneously, and study the implications of that violation for fairness metrics and methods. Real life frequently violates these assumptions as well, but often on different time scales; while people apply for a relatively few jobs over their lifetime, information access systems make millions of decisions about how to rank items over short time periods.

Information retrieval and recommender systems also have well-understood data sets and strong community norms of demonstrating performance on benchmark data sets. Many of these data sets are amenable to various forms of fairness analysis as well. This data availability may enable the study of fairness concerns over different, and sometimes much more, data than are widely available for other applications. The details will of course change when applying fairness research tested on these kinds of data to other applications, but information access may be a useful testing ground for mathematical or computational techniques with broader applicability.

Finally, information access represents a domain with substantial impact on lives, livelihoods, and perceptions of the world, but a very

different impact than the often-studied domains of criminal justice or lending. It may, therefore, present an opportunity to experiment with fairness measures, possibly even in real systems, where the human cost of getting it wrong or making the situation worse is quite different.

While great care is required to avoid abstraction traps when attempting to translate from fair information access to other problem settings (Selbst *et al.*, 2019), we think there is much that information access has to say with implications for other domains.

Information access has the potential to improve fairness research more broadly, as lessons learned studying information access may be applicable in additional domains considered by fairness researchers as well.

4.10 Navigating the Problem Space

Our goal with this section is to provide guidance to enable researchers, developers, students, and others building or affected by information access systems to consider potential harms, particularly harms related to fairness and discrimination, that can arise in information access. We do not claim this treatment is complete, and problems do not necessarily fall cleanly into one problem or another; scholars may also disagree with some of our categorizations here. Humanity is messy, and attempting to categorize the ways in which it can be harmed by technology is necessarily a messy and imprecise endeavor. We find this framing useful, however, for organizing our own understanding of the subject, and will use it to organize our discussion in the remaining. We submit that clearly describing and characterizing the harm(s) to be measured or mitigated in a particular work is more important than determining precisely which box it occupies in a rigid taxonomy, and we hope that our treatment provides a useful starting point for developing such clear descriptions.

In the rest of this monograph, we describe existing work and needed future research to address some of the harms cataloged in this section.

5

Consumer Fairness

Having described the problem space of fairness in information access, we now turn to surveying the literature to date on various definitions, methods, and metrics for fairness, beginning with consumer fairness. As noted in Section 4, information access systems are often in the position of mediating between providers of items or information and consumers of recommendations who are interested in those items. While fairness concerns may arise for any stakeholder in the system, these two groups have the most direct stake in the fairness properties of a recommender system and are the most widely-studied. Table 5.1 summarizes the key work we cite in this section.

Consumer fairness is concerned with how an information access system impacts consumers and sub-groups of consumers, and whether those effects are fair or result in unjust harms. For example, if a system is delivering recommended job postings to job seekers, it might be a fairness concern that different sub-groups of users, women, for example, could receive lower quality results than others.

Table 5.1: Summary of articles in consumer-side fairness.

Measuring group fairness	Yao and Huang (2017), Ekstrand *et al.* (2018b), and Mehrotra *et al.* (2017).
Enhancing group fairness	Yao and Huang (2017) and Abdollahpouri (2020).
Recommended Items	Nasr and Tschantz (2020), Kamishima and Akaho (2017), and Li *et al.* (2021).
Fair User Embeddings	Beutel *et al.* (2017), Edwards and Storkey (2016), and Madras *et al.* (2018).

5.1 Individual Fairness

The distinction between group and individual fairness is relevant for consumer fairness. As described in Section 3, individual fairness considers how individuals are treated by the system and whether similar users have similar experiences or quality of service within the system.

One of the most basic outcome measures that can be applied is the accuracy of results produced. Typical evaluation measures such as recall or nDCG as described in Section 2.5.2 can be used in offline experiments to determine the degree of accuracy that each user experiences in the system. While the central tendency of such measures form standard evaluation metrics for information access, the question of individual fairness calls for an examination of the distribution of utility across information requests, possibly marginalized to one dimension (such as user or query). In an extreme case, one might see a bimodal distribution of the evaluation metric, with some users getting accurate results and other quite inaccurate results. In such a case, the average performance is not capturing the user experience well; in particular, some users are being poorly served.

A form of evaluation that looks at the system's minimum performance would provide a form of corrective to this type of individual fairness problem. The "fairness without demographics" approach described by Hashimoto *et al.* (2018) works to solve this problem by constraining performance for all users within a particular error region so that the discrepancy in accuracy across users can be controlled overall. We are not aware of work applying this form of individual fairness in information access systems.

Constraining the system's accuracy distribution is a rough form of individual fairness with a kind of Rawlsian logic. It amounts to an assertion that all users are similar to all others, and thus can be used to ensure that all users are getting some basic level of service from the system. However, enforcing this type of fairness produces some challenges for information access systems, especially personalized ones. For example, User A with a small user profile (a *cold start* user) is generally expected to get less accurate recommendations than User B with a more extensive profile. It would not be a good solution to deliberately corrupt the recommendations for User B in order to equalize the accuracy of their results.

It is possible to take a more textured view of individual fairness in keeping with "similar users, similar results" rubric. For example, we could control for profile size in comparing accuracy distributions, ensuring that we only compare the system's performance for User A against other cold-start users. Thus, profile size itself might become a source of unfairness, and this could well be true of other features along which we might compare users to determine their similarity for the purposes of individual fairness. From this standpoint, it would be considered fair for fans of action movies to get similarly good results and devotees of documentaries to get similarly bad results from a recommender, an outcome that stretches what we might want a normative definition of fairness to provide.

5.2 Group Fairness through Disaggregated Utility

One major consumer-side group fairness problem is to determine whether the system provides comparable quality of service or utility to different groups of consumers, or whether there are groups — especially protected groups — whose information needs are systematically underserved by the system. One way to assess this is by performing the same kind of utility-based evaluation that is usually used to evaluate the system's effectiveness, such as an offline accuracy evaluation or an online A/B test, and disaggregating utility by consumer groups. That is, rather than computing an overall mean utility per user, computing average utility for each group.

In many cases, this utility can be operationalized through measures of the system's ability to meet information needs: click-throughs on search results or recommendations, ranking accuracy metrics such as nDCG or ERR, etc. Consumer fairness studied in this way does not bring anything new to the problem of evaluating the system, except how the results are broken down and analyzed. An overall metric $\bar{\mu} = \frac{1}{n}\sum_{\pi} \mu(\pi)$ is replaced by a per-group metric:

$$\mu_G = \frac{1}{n_G} \sum_{\pi, \rho : \rho_{\text{global}} \in G} \mu(\pi) \tag{5.1}$$

This metric can then be tested for significant between-group differences to assess whether some groups are experiencing better effectiveness (for whatever reason) than others. It applies to both online and offline measures of effectiveness.

Some applications, as noted above, may have further dimensions of utility connected to objective qualities of an item. For example, in a job opening recommender system, job listings have salaries. If protected group users receive on average lower-salary listings, this could be considered unfair regardless of other personalization considerations or equal satisfaction metrics, depending on the goals and context of the application.

Mehrotra *et al.* (2017) performed a disaggregation of user satisfaction with a search system across multiple measures (graded utility, page clicks, query reformulations, and successful clicks), finding differences between user age groups and genders; the system was more effective for older users than younger users across all measures. They further employed matching to control for query type and difficulty, to determine if differences in effectiveness were due to demographic differences in the queries issued; after context matching, both age and gender differences in satisfaction reduced almost to zero.

Yao and Huang (2017) focus on predicted rating fidelity, developing several unfairness metrics capturing different types of disparate prediction errors for protected and unprotected groups. They measure overall disparate error as well as separately analyzing over- and under-predictions: does the system systematically under- (or over-) estimate some users' preference more than others?

Ekstrand *et al.* (2018b) disaggregated offline top-N performance — as measured by nDCG — by age and (binary) gender for collaborative filtering algorithms trained on movie ratings (with the MovieLens 1M data set) and on music plays (with the Last.FM 1K and 360K data sets), finding statistically significant differences in utility between gender and (in some cases) age groups, although not always in the same direction. They further showed that this difference was not explainable by differences in user profile size, and that resampling training data to have equal gender representation had the effect of substantially reducing cross-group utility differences.

5.3 Disparate Effectiveness

The studies by Mehrotra *et al.* (2017) and Ekstrand *et al.* (2018b) represent different approaches to — and extents of — understanding the reasons for observed differences. Both begin with demonstrating the *existence* of a disparity in group outcomes: some groups receive better quality of service, as measured by the result quality metric. Such discrepancies bear similarity to disparate impact, in that there is a difference in outcomes for different groups, but differs in a crucial respect: unlike a typical classifier in the settings in which disparate impact is considered, an information system is not making decisions *about* the consumers that differ by group. We therefore refer to this kind of unfairness as **disparate effectiveness**: the system is more or less effective (in this case capable of satisfying information needs) for different groups of users. Identifying this disparate effectiveness is relatively straightforward. If there is a per-user or per-query measure of result utility, aggregating that over users' group membership and looking for (statistically significant) disparities detects the existence of disparate effectiveness.

That is only the beginning of studying fair utility, however. Disparate effectiveness can be caused by biases in any portion of the information access pipeline (Sections 2.1 and 4.7). Narrowing down potential causes is crucial for identifying whether and how to address disparate effectiveness, particularly since — as we will discuss shortly — regularizing it away is often is not a desirable strategy. There are several pathways that could give rise to disparate effectiveness:

- The outcome measure may not satisfy **measurement invariance** with respect to consumer groups: that is, users in two different groups with the same subjective experience of satisfaction of their information need may still respond to the system in different ways, such that a behavior-based measure of satisfaction (such as clicks or session length) may measure satisfaction differently for them.
- The system's ability to model document relevance may depend on the **availability of training data**, and thus the system is not able to learn as effectively how to meet information needs distinct to minority groups of users.
- **Item relevance** to the needs of different groups may differ in a way that the model may hold constant across all users or groups of users, so minority users are forced to use a relevance model optimized for the majority group.
- There may be **mediating factors** between an information need and its satisfaction that systematically differ between groups that in turn affect the system's ability to deliver satisfactory responses.

Mehrotra *et al.* (2017) address mediating factors through their context-matching design: by matching queries as closely as possible on multiple dimensions that affect their difficulty for the system, they are able to control for many of these factors. The fact that this control eliminated most of the disparate effectiveness is evidence that groups' differing satisfaction is mostly a result of these mediating factors. If a query has the same difficulty, groups tend to have the same satisfaction with the system's results for that query. This does *not* imply that observed differences are therefore not evidence of unfairness and do not need to be addressed; rather, it points to *where* the differences are. Younger users have lower satisfaction *because* they issue queries that are more difficult for the system to satisfy. Identifying the kinds of queries that present such difficulties, and are more frequently issued by under-served groups, provides a pointer to where engineering effort can be spent to improve quality of service for users currently receiving worse results.

The matching design is very powerful for isolating these effects, as it allows for variables known to affect query difficulty to be held as constant as possible between groups. It has the downside, however, of discarding a great deal of data (in this case, queries) that cannot be matched. If the disparate effectiveness arises primarily from those queries that cannot be matched, a matching design may obscure a real inequity in quality of service.

Ekstrand *et al.* (2018b) targeted one specific mediator, profile size, with a linear model, and found that it did *not* explain observed differences in recommendation quality. They also investigated the impact of availability of training data, and found that it did have significant impact on the observed differences. Downsampling is not necessarily an advisable approach in actual applications, because throwing away some of a group's data just because there is more of it is questionable machine learning engineering practice, but it is a useful strategy for narrowing down the causes of an observed discrepancy.

Neither of these approaches is strictly better than the other; they achieve different and complementary objectives through different means. Studying and ensuring consumer group fairness requires a variety of tools, and the field is still so new that systematic understanding of the strengths and weaknesses of different methodologies has not yet been developed.

5.4 Providing Fair Utility

Detecting and quantifying inequitable distributions of system utility is one thing; correcting them is another. Yao and Huang (2017) introduce regularizers to remove discrepancies in rating prediction errors; regularizers can also be employed to target various other discrepancies.

Examples of post-processing approaches take the form of re-ranking recommendations to improve their fairness properties. These approaches have typically focused on provider-side fairness, but Abdollahpouri (2020) considers user groups based on their level of interest in popular items and shows that, across recommendation algorithms, users with an interest in less-popular niche items were not receiving recommendations in line with their interests. Abdollahpouri also presented a re-ranking

approach based on the idea of calibration (Steck, 2018) to improve the fairness for these user groups.

Removing disparate effectiveness through an algorithmic intervention is not the obviously correct solution in many cases, however. Providing one user better results than another does not take anything away from the under-served user that they may otherwise have obtained and may not violate legal or ethical norms of fair treatment. This is different from classical fair decision making settings, where disparities such as a qualified borrower's chances of being approved for a loan differing on account of their race or religion are considered by many ethical and legal frameworks to be unacceptably discriminatory. It also differs from the provider fairness context, where giving one provider a recommendation slot *ipso facto* prevents another provider from occupying that slot and obtaining the benefits thereof. In particular, and in contrast to provider fairness, the well-served user's high-quality results are not usually the *reason* the under-served user receives worse results, and decreasing their result quality in the name of fairness is itself arguably unfair. Information access quality is not a rivalrous good, a fixed amount of which can be allocated across the users; we can improve experience for some users without hurting others at all. We therefore recommend caution when using regularizers or other algorithmic techniques designed to simply reduce disparaties in system consumer-side effectiveness.

Another approach is to treat the inequity through engineering *process*. If an analysis of mediating factors identifies that an under-served group issues queries that are systematically more difficult to satisfy in an identifiable way, prioritizing efforts to improve the system's ability to handle those queries, instead of efforts that will primarily improve quality for users already receiving the system's best results, can address the inequity (and may improve service for the majority group as well).

5.5 Fairness Beyond Accuracy

While much of the work on consumer fairness focuses on the quality of recommendations, some work looks at other aspects of consumer experience that may be discriminatory, such as stereotyping or the specific items users receive. Ali *et al.* (2019) studied the distribution of

ads on Facebook to understand potentially discriminatory impact in the visibility of different kinds of ads. They found that even when an advertiser wishes to have fair distribution of their ad, for example to ensure that an ad for a job opening is seen by people of all genders, the combination of relevance optimization and market dynamics results in disparate distribution of ads across racial and gender lines. Nasr and Tschantz (2020) describe bidding strategies to attempt mitigate such effects and ensure fair ad distribution even when the platform does not provide it.

More generally, Kamishima and Akaho (2017) presented a probabilistic test for the independence of results from a user's (or item's) protected class. Fairness, under their construct, is when the probability of a particular item being recommended is independent of the user's protected class. They then incorporated this idea into a loss function for a matrix factorization collaborative filtering algorithm to optimize the system to produce independent results. This can be useful in any context where users should not be recommended different types or sets of items on the basis of their group membership. Li *et al.* (2021) build on this independence objective in two ways: they allow fairness to be personalized, such that different users have different sensitive features they don't want affecting their recommendations; and they adopt a causal model to produce recommendations without causal pathways from the sensitive features to the recommendation lists.

Consumer fairness also extends beyond the items recommended, and can be applied to inner components of information. Beutel *et al.* (2017) present an approach to learning fair representations in a way that can be applied to consumers, by learning embeddings (such as the user and item embeddings in a recommender system) in an adversarial setting set up to minimize the ability to predict a user's sensitive attribute, such as gender, from their embedding. This has the potential to reduce stereotype effects in resulting recommendations, among other applications to both consumer- and provider-side fairness, although we have yet to see it deployed in this way. Similar ideas are explored by Edwards and Storkey (2016) and Madras *et al.* (2018).

5.6 More Complex Scenarios

Most of the work to date on consumer fairness assumes rather limited group fairness settings. In particular, it often assumes that only a single protected group, or a single dimension of sensitive attributes, needs to be considered for fairness; to the extent that they do consider multiple dimensions, these are considered separately (e.g. age and gender, but not combinations thereof). But a job recommender, for example, may need to meet simultaneously meet constraints having to do with race, gender, religion, and other types of protected categories, as determined by applicable laws and organizational requirements. As noted above, the complexities of the interaction of multiple protected categories have been explicated by Crenshaw (1989) and others under the framework of intersectionality. In the fair machine learning literature, it has been studied under the topic of *rich subgroup fairness* (Kearns *et al.*, 2019). In recommender systems, there is some research involving subgroup fairness across providers (Sonboli *et al.*, 2020b). However, no existing work addresses the compound nature of the disadvantage that Crenshaw highlights as characteristic for individuals who find themselves at the intersection of multiple protected identities.

Another simplification in this model of consumer-side fairness is that it assumes categories are binary (protected vs unprotected) rather than constellations of attributes, including continuous qualities. It is possible that some existing models could be extended to handle continuous sensitive features (age or income, for example) but there is not any extant work in recommendation fairness along these lines as of this writing.

6

Provider Fairness

We now turn to the second of the two primary stakeholder typically considered in multistakeholder information access analyses: provider fairness is concerned with fairness towards the providers (or *producers*) of the content or items the system makes available to its users. This primarily considers systems where content providers create and publish content that they want consumed, and for which they obtain some benefit from having users discover their content. This may be a direct tangible benefit, such as subscription, advertising, or pay-per-play revenue; it may be indirect, such as the reputational benefits that accrue to journalists or academics for producing widely-read content; or it may be intangible benefits, such as the satisfaction of providing useful content to readers. Under the definition of information access with which we opened Section 2, this aspect of fairness considers the impact of the system on providers who gain utility from the system satisfying a user's information need with an item they provided. Table 6.1 summarizes key papers we cite here.

These benefits are often abstracted under the notion of *exposure* (or *attention*) (Diaz *et al.*, 2020; Biega *et al.*, 2018): an item, and therefore its provider, appears in result lists, and users have the opportunity to

Table 6.1: Summary of articles in provider-side fairness.

Measuring group representation	Das and Lease (2019), Zehlike *et al.* (2017), Sapiezynski *et al.* (2019), Deldjoo *et al.* (2019), Yang and Stoyanovich (2017), Raj *et al.* (2020), Ekstrand and Kluver (2021), and Epps-Darling *et al.* (2020)
Enhancing group representation	Celis *et al.* (2018), Ekstrand and Kluver (2021), Zehlike *et al.* (2017), and García-Soriano and Bonchi (2021)
Measuring individual utility	Diaz *et al.* (2020) and Biega *et al.* (2018)
Measuring group utility	Biega *et al.* (2018), Diaz *et al.* (2020), and Singh and Joachims (2018)
Enhancing group utility	Biega *et al.* (2018), Diaz *et al.* (2020), Singh and Joachims (2018), Kamishima *et al.* (2018), Burke *et al.* (2018), and Zhu *et al.* (2021)
Pairwise fairness	Beutel *et al.* (2019) and Narasimhan *et al.* (2020)

interact with these items. We can treat result list opportunities as a resource, in which case the system is distributing these resources across the different providers (either individually or by groups), and we are concerned with whether or not that allocation is fair. For example, in the job candidate search scenario, when an employer is looking for people to hire, different protected groups (e.g., gender and ethnic groups) should be treated fairly in terms of their members appearing in recommended candidate lists. In music discovery, fairness would arguably require different artists whose work is equally relevant to a user's taste to have comparable exposure in their recommendations and streams.

In many scenarios, there are multiple parties who could be considered the provider of an item, at different levels and with different roles. For example, in a news portal, both individual journalists and publication venues are providers of a news article. In music recommendation, songs and albums have recording artists and record labels, as well as additional providers such as songwriters. Movies and television shows typically have a long list of contributors who could be considered providers. Studies of provider fairness typically focus on just one type of provider, but recognizing the diversity of provider relationships helps contextualize these concepts in the broader space of provider-side impacts of retrieval and recommendation.

Allocating utility is not the only way that an information access system can affect content providers, although it is the most commonly studied. In this section, we will also discuss problems and research related to other ways information access systems may unfairly harm (or help) providers.

Finally, we note that diversity is often closely related to provider fairness — indeed, one question we are often asked when presenting work on provider fairness is how this differs from diversity. As noted in Section 4.8, system modifications intended to enhance the diversity of results may be useful tools for improving the system's provider fairness, but diversity and fairness respond to different normative concerns and demand different metrics. Diversity in recommendation and search results is mainly focused on consumer intent, intending to present results that meet a wide range of users' topical needs. In contrast, provider fairness is motivated by justice concerns to ensure that different providers receive fair opportunity for their content or products to be discovered.

Our presentation here builds on the integration of provider fairness metrics for rankings provided by Raj and Ekstrand (2022); Table 6.2 summarizes the metrics we discuss. The literature to date differs in whether it establishes fairness *metrics* or fairness *constraints*; we here present the constructs in their original form, normalized for notation; constraints can often be converted to metrics, and Raj and Ekstrand provide metric versions of some of these constraints. Zehlike *et al.* (2022) and Kuhlman *et al.* (2021) provide additional comparisons and summaries of provider-fair ranking; in particular, Zehlike *et al.*'s treatment provides a particularly thorough discussion of the normative principles underlying different ranking metric decisions.

6.1 Provider Representation

Many constructs for provider fairness are concerned in some way with *representation*: are the providers of items returned representative of the broader population, or some other reference distribution of provider groups? It is always a group fairness construct, as this kind of representativeness is meaningful to the extent that item providers are

Table 6.2: Fair ranking constructs and their objectives. Adapted from Raj and Ekstrand (2022) by permission of the authors; metric names from that paper, except for Pair.

Metric(s) Goal	Section(s)
PreF_Δ (prefix fairness, Yang and Stoyanovich, 2017) *Each prefix representative of whole ranking*	
AWRF_Δ (attention-weighted rank fairness, Sapiezynski *et al.*, 2019) *Weighted representation matches population*	6.1.1
FAIR (Zehlike *et al.*, 2017) *Each prefix matches target distribution*	6.1.1
DP (demographic parity, Singh and Joachims, 2018) *Exposure equal across groups*	6.2.2
EUR (exposed utility ratio, Singh and Joachims, 2018, orig. DTR) *Exposure proportional to relevance*	6.2.2
RUR (realized utility ratio, Singh and Joachims, 2018, orig. DIR) *Discounted gain proportional to relevance*	6.2.2
IAA (inequity of amortized attention, Biega *et al.*, 2018) *Exposure proportional to predicted relevance*	6.2.1, 6.2.2
EEL, EER (expected exposure {loss, relevance}, Diaz *et al.*, 2020) *Exposure matches ideal (from relevance)*	6.2.1, 6.2.2
EED (expected exposure disparity, Diaz *et al.*, 2020) *Exposure well-distributed*	6.2.1, 6.2.2
Pair (Beutel *et al.*, 2017; Narasimhan *et al.*, 2020) *Pairwise rank accuracy equal across groups*	6.3

representative of their groups. Unfortunately representation is an overloaded term; we are not concerned here with the internal or external representations of any individual provider, but rather with how the system represents the *space* of providers to the user.

This kind of provider fairness can be concerned with a representational harm, in that users who experience a skewed view of the population of item providers may develop (or have reinforced) an imbalanced view of who creates content; or it may be a proxy for a distributional harm, as result lists in which particular provider groups are systematically under-represented are likely to result in unjust denial of exposure or utility to those groups.

These fairness constructs are usually independent of relevance: it is assumed that the lists in question are already optimized for utility, and their fairness is measured as a separate concern.

6.1.1 Measuring Representation

Provider group representation in result lists is typically operationalized through a distribution over provider groups. There are therefore three

components to a representation-based measurement of information access results:

- A multinomial target distribution P_{target} over provider groups $\mathcal{G}$
- A distance function Δ that computes the distance between two distributions over provider groups
- A means of computing group distributions P_π from the list and comparing them to the target distribution

When there are only two provider groups under consideration, such as a protected and unprotected group, the distributions reduce to binomials.

The simplest form of measuring provider representation is to compute a multinomial from a system decision π based on the number of times each item appears (Das and Lease, 2019); for a single group $G \in \mathcal{G}$, this is:

$$P_\pi(G) \propto |\{d \in \pi : p_d \in G\}|$$

Unfairness can then be defined using the distance $\Delta(P_\pi, P_{\text{target}})$ (e.g., the Kullback-Leibler divergence, $\Delta(P_\pi, P_{\text{target}}) = \Delta_{\text{KL}}(P_\pi \| P_{\text{target}})$); a normalized (un)fairness metric can be computed by minimax-scaling the divergence (Das and Lease, 2019). Simple binomial fractions are also the family of metrics for which Kırnap *et al.* (2021) developed sampling procedures for estimating with incomplete labels.

This approach is good for measuring overall list composition, but it does not take into account the relative visibility of different positions in the ranking. Even when we are concerned with representation, not the distribution of utility, it is reasonable to expect the distribution among items at the top of the ranking to have a larger impact on the users' perception of the space than items further down the list.

There are two basic approaches in the literature to date for incorporating rank position into a representation-based fairness construct. The first is to consider *prefixes* of the ranking. Zehlike *et al.* (2017)

presented a prefix-based approach for binomial group fairness that applies hypothesis tests to prefixes of the list of increasing length. For each prefix $\pi_{\leqslant k}$ (of length k) with m items from the protected group, they compute the probability that a list of that length would have at most m protected-group items under the null hypothesis that its item providers were independently drawn from the binomial target distribution ($F_{\text{target}}(m|k)$, where F_{target} is the cumulative distribution function of the target distribution). If the null hypothesis is rejected (i.e. $F_{\text{target}}(m|k) < \alpha$) for a prefix of the ranking, after correcting for multiple comparisons, the ranking is deemed to be unfair. This ensures that a ranking cannot be considered fair unless provider groups are evenly represented throughout the ranking.

Another approach is to employ a discount factor $\delta(r)$ to down-weight representation further down the list. Sapiezynski *et al.* (2019) do this, so that:

$$P_\pi(G) \propto \sum_r \delta(r) \mathbb{I}(p_{(\pi_r)} \in G) \tag{6.1}$$

where $\mathbb{I}(\cdot)$ is the $0, 1$ indicator function. This construct, called "Attention-Weighted Rank Fairness" (AWRF) by Raj and Ekstrand, supports more than two groups, and can be extended to real-valued or mixed group membership weights $w(d, G)$, either to represent multiple group membership or uncertainty about the group alignment of an item. If δ forms a distribution such that $\delta(r)$ is the probability of the user selecting the item at position r, then the distributions computed under this definition are the probability that the user will select an item provided by a member of a particular group.

As before, the distribution from the ranking can be compared to the target distribution using a suitable discount function; this can be Kullback-Leibler divergence (Das and Lease, 2019), cross-entropy (Deldjoo *et al.*, 2019), or another suitable difference; in their study, Sapiezynski *et al.* (2019) used the Z-approximation for the binomial test statistic.

So far, we have not discussed P_{target}: how do we determine the ideal target to which a ranking's group distribution should be compared?

Each of these approaches abstracts over this target; Sapiezynski *et al.* (2019) call it the *population estimator*, assuming that the goal is for the providers in a ranking to be representative of the broader population from which they are drawn; Deldjoo *et al.* (2019) call it the *fair distribution*, assuming we have some target we deem "fair". The precise choice of target distribution will depend on the domain, application, and specific fairness goals. Potential reasonable choices include:

- Uniform (Deldjoo *et al.*, 2019)
- The overall population of item providers
- The set of providers of items at least marginally relevant to the information need (Yang and Stoyanovich, 2017)
- An estimate of the distribution in society at large

Calibrated fairness (Steck, 2018) compares result lists to the user's past activity: under this definition, the group distribution in the user's reading, listening, or purchasing activity is used as P_{target}. For example, this would consider a music recommender to be gender-fair if the mix of artist genders in each user's recommendations match the mix in their previous listening history.

There is not currently consensus, or even much study, of the relative strengths and weaknesses of these approaches. Raj and Ekstrand (2022) and Kuhlman *et al.* (2021) provide some direct comparisons, but papers typically measure a particular fairness construct without comparing with other metrics (aside from possibly variants on the same idea). This isn't as bad as it seems, however, because construct validity — measuring the intended fairness objective — is a more important property for a fairness metric than consistency with prior results.

6.1.2 Studies of Gender Representation

Representation-oriented metrics have formed the backbone for studies of gender fairness in recommender systems that are focused primarily on understanding system behavior, not on developing fairness constructs.

Both of these studies operationalize gender fairness as the fraction of recommendations or interactions that are with items provided by women ("% Female").

Ekstrand and Kluver (2021)[1] studied this in the context of book recommendation, looking at gender representation in repositories, user reading or rating histories, and collaborative filtering recommendations across multiple book recommendation data sets. Their work documents a large composite data set for studying fairness in book recommendation[2], that is likely useful for many more search and recommendation studies, particularly ones looking at provider fairness.

Ekstrand and Kluver found that women were better-represented among the authors of books read or rated by users than they were in the Library of Congress catalog, and that users tended to read more men than women but were highly diffuse in their gender tendencies. They employed a hierarchical Bayesian model to account for different user activity levels and produce smoothed estimates of users' gender biases, which they then used in a regression to examine whether collaborative filtering recommendation lists had gender balances that correlated with users' reading histories (*calibrated fairness*). They found that collaborative filters did reflect users' biases with respect to author gender in their recommendation lists, although to different degrees.

Epps-Darling *et al.* (2020) conducted a similar analysis of music listening activity on Spotify with respect to artist gender. They found that male artists dominated streaming activity in both recommender-generated ("programmed") and user-generated ("non-programmed") activity; they also found, though, that increased prevalence of female artists in programmed streams was correlated with increased non-programmed listening activity for female artists. They also looked for a difference in listening to women artists along user gender and age, but did not find an user demographic differences in the share of streaming activity that went to female artists.

[1] An earlier version is provided by Ekstrand *et al.* (2018c).
[2] https://bookdata.piret.info

In contrast to this lack of an interaction effect between artist gender and user demographics in music listening, Thelwall (2019) found in his analysis of GoodReads reviews and ratings that users are more likely to give high ratings to authors of their own gender.

While much of the work on provider-side fairness is concerned with defining metrics and optimization strategies, these studies provide more extended examples of studying the sources of bias and the propagation of such biases through standard recommendation algorithms.

6.1.3 Ensuring Fair Representation

Methods for ensuring fair representation often follow from the metrics that implement a fairness construct. One common way to provide representational group fairness is through re-ranking. In binary-group settings, a greedy approach that selects the best item from the original ranking that does not violate the fairness constraint (Celis *et al.*, 2018) or make representation worse (Ekstrand and Kluver, 2021) can be effective. Zehlike *et al.* (2017) greedily process the list, picking the best item available (by the ranking's underlying relevance scores) if it would not make the protected group under-represented, and selecting the best protected group item if it is necessary to prevent under-representation.

These approaches, properly implemented, maintain *in-group monotonicity* (Zehlike *et al.*, 2017): the order between items within a particular group is preserved in the fairness-enhanced re-ranking, and items are only reordered with respect to items from other groups. Zehlike *et al.* further prove that their greedy approach results in the ranking with maximal overall utility subject to the binomial fairness constraint and in-group monotonicity, assuming the accuracy of the system's underlying utility estimates. Ekstrand and Kluver (2021) show empirically that greedy approaches need not result in substantial loss on utility-based evaluation metrics, at least in their experimental setting; Gómez *et al.* (2021) provide a similar result for geographic representation in MOOC course recommendations.

One concern often raised about group representation fairness is that majority-group providers of relevant content may be moved aside in order make room for the items needed to achieve group fairness

goals. Under some theories of equity, such as anti-subordination, this is expected and acceptable. García-Soriano and Bonchi (2021) address this concern by using randomness to ensure that it is not always the same items that are bumped aside, but that an individual fairness bound is preserved while meeting representative group fairness objectives. Exposure, discussed in the next section, provides another perspective on relating group and individual fairness.

6.2 Provider Exposure and Utility

As noted at the beginning of this section, many provider fairness constructs are designed to ensure that providers have fair opportunity to realize the utility that arises from providing content responsive to users' information needs. Even representational measures of provider fairness are often intended as a proxy for access to utility (see e.g. Ekstrand *et al.* (2018c) and Sapiezynski *et al.* (2019)).

A more recent line of fair ranking constructs shifts this discussion in four important ways:

- Assuming that measures of relevance produced by an information access system are good proxies for the value of an item to a user, such that the inclusion of a high-scoring item is worth more to the provider, as well as to the user.
- Directly measuring **exposure** (or *attention*) as a resource that the system should distribute fairly among providers.
- Relating provider-side utility, abstracted through exposure, to consumer-side utility.
- Measuring fairness over repeated or stochastic rankings, rather than a fixed ranking in response to a single information need.

The first of these changes involves an aspect of the WYSIWYG assumption, namely that users' preferences, as filtered through the information access system and output as predicted utility or preference, are unbiased indicators of the value of a item. As opposed to the prior construct of representation, which assumes all list appearances have

utility, the exposure construct considers utility to be a function of the match between user and item, as the system predicts it. This avoids one of the drawbacks of a purely representational approach: that the fairness metric can be satisfied with the inclusion of irrelevant protected group items, which are unlikely to attract user interest.

The second and fourth of these changes connect with the idea of browsing models used to evaluate information access systems (Section 2.5.2), and with common patterns in the presentation of result lists. Because rankings as actually presented to users are often short, a single list contains only a small number of opportunities for exposure. Further, because users are more likely to engage with items at the top of the ranking than the bottom, these slots are not of equal value: the first-ranked position (under most browsing models) provides more exposure to its occupant than the third or seventh position.

Therefore, provider utility is typically measured cumulatively (or in some cases amortized) over a sequence of recommendation results delivered to users, or as the expectation over a distribution of rankings defined by a stochastic ranking policy. In important ways, it is often difficult — if not impossible — to fairly allocate exposure in a single ranking. Considering fairness over sequences or distributions allows for a rich family of fairness constructs that are still achievable, at least in approximation. Note however that this kind of evaluation is only appropriate if aggregate utility over time is an appropriate scoring mechanism. Some fairness contexts, however, might still require that each generated ranking be fair with respect to protected groups: lists of job candidates in recruitment context are an example.

6.2.1 Individually-Fair Exposure

As noted in Section 3.2.2, the key idea of individual fairness is that similar individuals — in this case, providers of items — should be treated similarly (Dwork *et al.*, 2012); in exposure-oriented analyses of provider impacts of information access, that looks like receiving similar (opportunity for) exposure. Information access's focus on utility or relevance to an information need provides a relatively natural basis for assessing similarity with respect to the task: two items are similar if

they have similar relevance to the information need (either assessed by ground-truth relevance judgments or estimated by the system's relevance model). Individual provider-side fairness is often computed at the item level, ensuring that items are treated fairly without aggregating to the provider level; in practice there is little difference between these concepts.

Diaz *et al.* (2020) operationalize this by taking the *expected exposure* over a stochastic ranking policy $\boldsymbol{\pi}$. Defining exposure based on a discount model δ, so that exposure $\eta(d|\pi) = \delta(\pi_d^{-1})$, the expected exposure for a item is:

$$\mathrm{EE}(d|\boldsymbol{\pi}) = \mathrm{E}_{\boldsymbol{\pi}}\left[\eta(d|\pi)\right] = \sum_{\pi} \eta(d|\pi)\mathrm{P}_{\boldsymbol{\pi}}(\pi) \tag{6.2}$$

This exposure can then be compared to the exposure under a *target policy* $\boldsymbol{\pi}^{\text{target}}$; Diaz *et al.* used a policy that is uniform over all rankings that respect the relative relevance of documents to the information need as the target, and provide closed-form solutions for target exposure under two browsing models. With a target policy, we can compute the *expected exposure loss* as squared difference between actual and target exposure, over all documents:

$$\mathrm{EEL}(\boldsymbol{\pi}) = \sum_{d} \left(\mathrm{EE}(d|\boldsymbol{\pi}) - \mathrm{EE}(d|\boldsymbol{\pi}^{\text{target}})\right)^2 \tag{6.3}$$

If the policy $\boldsymbol{\pi}$ distributes exposure comparably to the target policy, then the difference in exposure under the two different policies will be low, and thus the overall squared difference will be low. This metric embodies the "equal exposure" principle: a fair ranking (policy) is one in which exposure is equally distributed among relevant documents. Item exposure can be converted to provider exposure by aggregating over a provider's items. Squaring just the expected exposure under the system policy yields *expected exposure disparity* (EED), a measure of how equally exposure is distributed among documents regardless of their relevance, a measure similar to Sapiezynski's discounted metric.

Biega *et al.* (2018) similarly relate exposure to consumer-side utility by requiring a provider's exposure to be proportional to their utility,

amortized over a sequence of rankings that may be in response to different queries:

$$\mathcal{A}(d) = \sum_{\pi} \delta(\pi_d^{-1})$$
$$\mathcal{R}(d) = \sum_{\pi,\rho} u(d|\rho)$$

Equitable attention is satisfied when $\mathcal{A}(d)/\mathcal{R}(d) = c$ for all items d; Biega *et al.* quantified violations of this principle through the L_1 norm $\text{IneqAttn} = \sum_d |\mathcal{A}(d) - \mathcal{R}(d)|$.

One source of complexity in computing these metrics for a whole system, responding to multiple information requests, is determining how to aggregate over those requests. Biega *et al.* (2018) take the sum over all rankings, regardless of query, and do this sum *before* comparing attention to relevance. This results in a measure of the overall attention a document or provider receives from the system, taking into account the relative popularity of various information needs, which is useful for approximating provider utility when the goal is to ensure that providers obtain fair renumeration (e.g. ad clicks) for their production work. Because attention and relevance are aggregated separately, however, a system can be fair by providing the correct exposure to items, but exposing them on the wrong queries.

Diaz *et al.* (2020) go the other direction, and compute the metric over stochastic rankings in response to a single (likely repeated) information request. This can be averaged over information requests, either with or without traffic-weighting (weighting a request by its relative frequency in the system logs). Comparing actual and target exposure on a per-request basis binds an item's exposure to the information needs for which it is relevant, so the system cannot achieve fairness by exposing content in response to the wrong requests. That feature, however, makes this measure difficult to apply in a recommendation context where users' information needs are assumed to be personalization and more or less unique.

Neither of these metrics have a naturally-interpretable scale, and are not suitable for comparing across data sets or experimental settings; they are only effective for comparing the fairness of multiple systems on the same sequence or distribution of information requests.

6.2.2 Group-Fair Exposure

These exposure concepts can be extended to group fairness by aggregating exposure over groups. Both Biega *et al.* (2018) and Diaz *et al.* (2020) describe group-based aggregations of their amortized attention and expected exposure metrics; as presented, these consist of aggregating attention and relevance (for amortized attention) or exposure (for expected exposure) by provider group before computing the loss metric:

$$\begin{aligned}
\mathrm{EE}(G|\boldsymbol{\pi}) &= \sum_{d:p_d \in G} \mathrm{EE}(d|\boldsymbol{\pi}) \\
\mathcal{A}_G &= \sum_{d:p_d \in G} \sum_{\pi} \delta(\pi_d^{-1}) \\
\mathcal{R}_G &= \sum_{d:p_d \in G} \sum_{\pi,\rho} u(d|\rho)
\end{aligned}$$

Unfairness can the be computed with the squared difference in groupwise exposure between system and target exposure, or absolute difference between group exposure and relevance.

Singh and Joachims (2018) propose parity constraints and ratio-based metrics for fair exposure with respect to binary protected groups under stochastic rankings:

$$\mathrm{EE}(G^+|\boldsymbol{\pi}) = \mathrm{EE}(G^-|\boldsymbol{\pi}) \qquad \text{demographic parity} \qquad (6.4)$$

$$\mathrm{EUR}(\boldsymbol{\pi}) = \frac{\mathrm{EE}(G^+|\boldsymbol{\pi})/\mathcal{R}_{G^+}}{\mathrm{EE}(G^-|\boldsymbol{\pi})/\mathcal{R}_{G^-}} \qquad \text{exposed utility ratio} \qquad (6.5)$$

$$\mathrm{RUR}(\boldsymbol{\pi}) = \frac{\mathrm{E}_{\boldsymbol{\pi}}\left[\mu(G^+|\pi)\right]/\mathcal{R}_{G^+}}{\mathrm{E}_{\boldsymbol{\pi}}\left[\mu(G^-|\pi)\right]/\mathcal{R}_{G^-}} \qquad \text{realized utility ratio} \qquad (6.6)$$

Demographic parity is a straight statistical parity constraint that ignores relevance and simply requires equal exposure; achieving this is

equivalent to minimizing groupwise $\text{EED}_{\mathcal{G}} = \sum_{G \in \mathcal{G}} \text{EE}(G|\boldsymbol{\pi})^2$. "Exposed utility ratio" and "realized utility ratio"[3] realize a similar logic as equal exposure or equitable amortized attention: the system is fair if each group gets the exposure it merits by producing content relevant to users' information needs. EUR is a direct ratio-based analogue to these metrics, while RUR incorporates the actual utility to the user in the numerator in addition to the overall utility in $\mathcal{R}_G$, through the use of the evaluation metric (a group-wise evaluation metric is the average of the metric value for documents provided by the group). Singh and Joachims motivated this as an offline approximation of click-through rate, so that this metric is closer to the measuring the distribution of actual user *engagement* instead of just exposure to users that *may* lead to engagement.

These metrics all implement variants on the groupwise analog of the equal expected exposure principle: a system is provider group-fair if it distributes exposure to provider groups commensurate with their utility with respect to the users' information needs. But as defined so far, they all have the drawback that they aggregate over groups *before* computing whether the exposure is merited for a particular information request or not. This is similar to the problem with aggregating an item's exposure and relevance separately across information requests and comparing total exposure to total relevance: not only may the system be able to achieve a good fairness score by exposing items to the wrong information requests, in group fairness it can be achieved by exposing the wrong items for a group. So long as a group has some relevant items, and some items are exposed, there is nothing in the fairness metric (except for RUR) that requires that the relevant items are the ones exposed. It can be achieved by randomly selecting items in a group-fair way with no attention to actual utility. Combining it with a utility metric in a multi-objective analysis would help, but is a step backwards from the promise of exposure-based metrics to integrate fairness and utility directly.

[3] We use here the names provided by Raj and Ekstrand (2022), as we believe they better reflect the general use of the terms disparate treatment and disparate impact than the original names of "disparate treatment ratio" and "disparate impact ratio" used by Singh and Joachims (2018).

One way to address this problem is to compute over- or under-exposure ($\text{EE}(d|\boldsymbol{\pi}) - \text{EE}(d|\boldsymbol{\pi}^{\text{target}})$) on a item-by-item basis, and then take groupwise aggregates of this exposure difference. The resulting metric will consider a system to be group-fair if no group's items are systematically under- or over-exposed more than another group's; this method has been adopted by the TREC 2022 Fair Ranking track.

6.2.3 Ensuring Fair Exposure

It is not sufficient to simply measure violations of fair exposure objectives; we often want to modify the system to provide results with greater fairness in their exposure. We only outline the approaches here, referring the reader to the individual papers for details.

As with representational fairness constructs, re-ranking can be a promising approach. Biega *et al.* (2018) describe a reranking stratgegy based on integer linear programming to ensure individual fairness of amortized attention. Gómez *et al.* (2021) re-rank algorithms using minimal-cost swaps to reduce unfairness in both exposure and representation. Given a stochastic policy $\boldsymbol{\pi}$ represented as a doubly-stochastic matrix, Singh and Joachims (2018) present a linear programming solution to produce stochastic rankings that satisfy their group fairness constraints.

Diaz *et al.* (2020) directly use expected exposure loss as a learning-to-rank objective. Singh and Joachims (2019) similarly adopted a fair policy learning framework to learn stochastic ranking policies that fairly allocate exposure. Their approach augments a standard utility maximization approach, that ensures the most relevant items receive the most exposure, with a lower-bounding inequality so that, for $u(d_i) > u(d_j)$, $\frac{\text{EE}(d_i)}{u(d_i)} \leqslant \frac{\text{EE}(d_j)}{u(d_j)}$. This ensures that while more relevant items get more exposure, the disparity in exposure doesn't outrun the difference in utility, thus addressing one of the drawbacks to meritocratic fairness (Joseph *et al.*, 2018) in which fairness can be achieved by giving the most relevant document all the exposure.

Kamishima *et al.* (2018) present yet another approach to providing fair exposure, at least in a binary sense, to providers from different groups. The provider-side element of their work formulates fair rec-

ommendation through statistical independence: $\mathrm{P}(d \in \pi_{\leqslant k} | p_d \in G) = \mathrm{P}(d \in \pi_{\leqslant k})$. They then regularize the recommendation model to penalize violations of this independence objective.

Finally, Burke *et al.* (2018) provide a more indirect approach that modifies neighborhood-based recommendation to ensure that neighborhoods are balanced between protected and unprotected groups, so that protected-group items have a good chance at being recommended.

Most techniques for provider fairness, including those described so far, focus on providing fair exposure in response to an established information need match: available query and document text for a search application or the normal steady-state case for recommendation. Zhu *et al.* (2021) examine provider fairnesss in *cold-start* recommendation (recommending new items that do not yet have sufficient user interactions for typical collaborative filtering approaches to recommend them), and adjust the cold-start process to maximize the minimal exposure (expressed by discounted cumultative gain) of each new item to ensure that the system is fair to new items from different providers or provider groups.

6.3 Fair Accuracy and Pairwise Fairness

Another way of conceptualizing provider fairness, that has very similar motivations to fair exposure but results in very different metrics, is to look at *pairwise accuracy* as a basis for fairness. As noted in Section 2.6.4, pairwise rank loss has long been used as a learning-to-rank objective for recommender systems (Rendle *et al.*, 2009).

Beutel *et al.* (2019) and Narasimhan *et al.* (2020) define fairness metrics based on pairwise accuracy. The key principle of these metrics is that a system is fair if it is not systematically more effective at correctly ordering relevant items from one group than it is from another — that is, the probability that d_+ will be ranked above d_- is conditionally independent of provider group given that d_+ has higher utility than d_-:

$$\mathrm{P}(d_+ \succ_\pi d_- | u(d_+) > u(d_-); p_{d_+} \in G) = \mathrm{P}(d_+ \succ_\pi d_- | u(d_+) > u(d_-))$$

In the case of a binary protected group, this can be further refined into *intra-group* and *inter-group* pairwise accuracy (Beutel *et al.*, 2019). Intra-group requires that the protected and unprotected groups have the same pairwise accuracy for ordering items within the group (or, equivalently, each group has the same ROC AUC):

$$\begin{aligned}\mathrm{P}(d_+ >_\pi d_- | u(d_+) > u(d_-); p_{d_+}, p_{d_-} \in G^+) \\ = \mathrm{P}(d_+ >_\pi d_- | u(d_+) > u(d_-); p_{d_+}, p_{d_-} \in G^+)\end{aligned}$$

Satisfying this constraint ensures that the system is not more accurate at modeling relative preference for items created by one group than another; for age discrimination in job candidate search, for example, it would ensure that the system is not systematically more accurate at estimating the relative qualification of older candidates than younger ones.

Inter-group fairness requires that the groups have the same pairwise accuracy when compared with an item of the other group:

$$\begin{aligned}\mathrm{P}(d_+ >_\pi d_- | u(d_+) > u(d_-); p_{d_+} \in G^+, p_{d_-} \in G^-) \\ = \mathrm{P}(d_+ >_\pi d_- | u(d_+) > u(d_-); p_{d_+} \in G^+, p_{d_-} \in G^-)\end{aligned}$$

Beutel *et al.* (2019) further extended these to leverage two-stage relevance feedback (e.g. clicking on an item, followed by a post-click signal of utility such as rating) to avoid simply optimizing to amplify click probabilities (a common signal for pairwise rank loss), and showed that group pairwise accuracy can be used as a regularization for a pairwise learning-to-rank algorithm like BPR (Rendle *et al.*, 2009) to penalize group disparities in ranking accuracy.

Pairwise accuracy can be estimated by sampling and only requires group and utility data for the sampled items, as opposed to the exposure-based metrics which — in their original form — require data across the complete ranking. This may make them more sample-efficient and/or easier to apply in partial data scenarios, but this potential benefit has not yet been well-explored.

Cui *et al.* (2021) present a similar mechanism, re-ranking result lists to preserve within-group ordering and optimize AUC while balancing fairness and accuracy loss.

6.4 Related Problem: Subject Fairness

As we noted in Section 4.3.3, subject fairness — treating the subjects of items, such as the subjects of news articles or the population in a medical study — has much in common with provider fairness, at least in terms of its structure. Since subjects and providers are both entities associated with items, many of the same metrics and fairness mechanisms can be employed; the only change needed is the item attribute considered. Subject fairness is also closely related to diversity, even more so than provider fairness, as it aims to ensure that the results contain representation from a wide array of possible subjects.

However, depending on the information access context, subject fairness may require revisiting a key assumption behind many of the metrics discussed above, namely the assumption of cumulative utility. It may not be sufficient for subject fairness to allow fair results at time $t + 1$ to compensate for unfair results at time t. For example, if we are concerned that image search results for "CEO" gives an unfair representation of the percentage of women in that position, we might not find it acceptable to mix 100% male result lists with the occasional over-representative female list. These lists go to different users and therefore do not avoid the representational harm we are seeking to avoid. In such a case, minimum (or at least distributional) properties of list-wise metrics will be of interest rather than (or in addition to) averages over many results.

Subject fairness introduces the challenge, also, of identifying the subjects. While items are often annotated with their creators, they are not always annotated with the relevant aspects of their subjects, at least in a machine-readable manner. More advanced content analysis techniques or extensive human annotation may be necessary to obtain the labels needed to pursue subject fairness, particularly for ensuring fairness to subject groups.

One sub-area of search that has seen significant work on subject fairness is in image search: Kay *et al.* (2015) and Metaxa *et al.* (2021) provide measurements and empirical techniques of gender and race biases by comparing representation in image search with estimated representation from the US Bureau of Labor Statistics; Singh *et al.*

(2020) provide another treatment. Otterbacher *et al.* (2017) build on this with evaluation and design recommendations for improving such systems and their human effects. Karako and Manggala (2018) provide a techniqe based on maximum marginal relevance (MMR), a diversification technique, for improving the diversity of a set of image results; Celis and Keswani (2020) apply MMR and propose a new technique that does away with MMR's need to compare results with each other (improving the ranking efficiency). Further, Celis and Keswani use textual descriptions from off-the-shelf image summarization algorithms to improve diversity and subject fairness without needing explicit image labels.

Outside of image search, subject fairness is not extensively studied in the fair recsys and IR literature. Rekabsaz *et al.* (2021) provide one example, approaching subject fairness in retrieved text passages by measuring whether each retrieved document is neutral or unbalanced in its presentation of sensitive groups, and prioritizing the retrieval of neutral documents (ones that either do not include sensitive group information or are balanced in their representation of it, as determined by the relative frequency of group-related keywords).

Subject fairness is also implicated in many examples of search-related harms, such as the representational harms towards Black girls documented by Noble (2018), but it has not yet received as much research attention — that we are aware of — in the research literature.

7

Dynamic Fairness

In Sections 5 and 6, we have considered fairness for consumers and providers (and subjects) at a *single point in time*: the current state of the system and its models should fare. While stochastic policies act over time, the treatment in Section 6.2 does not consider updates to the policy or changes to items or users.

Information access systems, however, operate in an iterated, changing environment. They continuously make new decisions, gain fresh users, and lose established users. This dynamism is particularly salient in an application with high item churn such as news recommendation, where articles may be superseded by fresher stories in quick succession (Karimi *et al.*, 2018). But even when items have longer lifetimes, as in music, items and providers will come and go from the system over time. In addition, seasonal changes and longer-term trends means that historical profiles of users may lose utility over time. For example, a user who searches for an entry level job at one point in time may be looking for a different kind of position in the future.

These dynamics are central to lines of research in recommender systems that consider the temporal aspects of markets and of user behavior (Jambor *et al.*, 2012; Harman *et al.*, 2014; Campos *et al.*,

2014; Basilico and Raimond, 2017; Zhang *et al.*, 2020). Note that this area of research is distinct from work that treats the problem of recommendation as one that involves temporally-extended and dynamic learning behavior, as in multi-armed bandit or reinforcement learning formulations. In these settings, the recommender system changes its policy over time in a process of exploring user preferences and item qualities, but for the most part, the items and users are considered a static aspect of the environment over which learning takes place.

Understanding the dynamics of information access systems in general is a relatively recent project (although there are historical examples, such as that of Fleder and Hosanagar (2009)), and ML fairness research is also only beginning to scratch the surface of dynamic fairness (D'Amour *et al.*, 2020). The need to study information access fairness over time is clear, but there is so far relatively little work on it. In this section we provide pointers that researchers wishing to explore this vital topic.

7.1 Feedback Loops

The dynamics of recommendation contexts have also been considered in the context of recommender system fairness. One of the most troubling aspects of algorithmic bias generally is the potential for destructive positive feedback loops within the system (O'Neil, 2017). Credit redlining provides an example. If a particular geographic area is determined to be too risky for lending, not only are current applicants impacted, but future ones as well. The system will not gather counterexamples that would help it identify the borrowers within the region that are actually good risks.

Hashimoto *et al.* (2018) study feedback loops in production systems from a fairness perspective. The authors model a population of users iteratively engaging with a system that trains using behavioral data. The model and supporting experiments in the context of predictive typing demonstrate that, over time, machine learning algorithms pay more attention to dominant subgroups of users as they lose underrepresented subgroups of users. The authors propose applying techniques from distributionally robust optimization to achieve more balanced performance, resulting in broad user retention. Zhang *et al.* (2019)

extend this work by analyzing the dynamics of fairness in sequential decision-making. Finally, in the context of predictive policing, Ensign *et al.* (2018) theoretically demonstrate how to filter feedback to improve the fairness of decision-making systems learning from feedback loops.

A well-known effect in information access systems is that of various presentation-related biases (Joachims *et al.*, 2017; Yue *et al.*, 2010). User are more likely to experience and rate items that the system itself suggests, and their interaction may be affected by where and how it presents those results. This is by design: the system is presenting items that it regards as ones users will want to interact with. However, this bias can cause a form of positive feedback, in which presented items gain in popularity, leading to greater bias towards presenting them, at the expense of other items (Chaney *et al.*, 2018; Fleder and Hosanagar, 2009). Positive feedback loops are inherently antithetical to fairness: they magnify small initial differences between item rating frequency into large ones as time goes on. It is also very difficult for new entrants to break into a market with positive feedback effects since they would have to gain traction against well-entrenched competition. Recommender systems therefore tend naturally towards unfairness, a tendency that a fairness-aware recommender system will need to continuously counter.

Feedback loops in recommender systems have been studied in a number of recent works. Chaney *et al.* (2018) examined the homogenization of recommendations in iterative environments. They find that recommendation systems, especially those based on machine learning, increase the consistency in recommendations across different users but also tend to increase the inequity of exposure across items. This phenomenon was termed "bias amplification" in work by Mansoury *et al.* (2020). Similar effects were found an information retrieval context in (Sun *et al.*, 2018). Multi-agent simulation techniques were used to provide a theoretical basis for such findings in Jiang *et al.* (2019). Some work in online learning contexts looks to rectify these biases and prevent bias amplification through the feedback loop; for example, Morik *et al.* (2020) use separate fairness and utility estimators to improve group fairness in dynamic learning-to-rank settings.

7.2 Dynamic Evaluation

In its most basic form, accounting for recommender system dynamics means moving away from a recommendation experimentation model that has the form of batch training followed by batch testing. Instead we need to incorporate the cycle of user arrival, recommendation generation, user response, and periodic system re-training. Off-line evaluation takes on the character of simulation of a recommender's evolution over time.

This type of evaluation has become standard in recommendation approaches that make use of reinforcement learning, in which the whole point is to develop algorithms that are compatible with a dynamic environment. Li *et al.* (2010) and Zheng *et al.* (2018) provide for more details about this algorithmic approach. By necessity, such training requires the application of *off-policy evaluation* because ground-truth user responses will only be available for a small subset of the recommendations that could generated and yet we need to model these responses as training input.

In practical deployments where fairness is a concern, the appropriate form of evaluation might be to consider the system's fairness properties over some particular time interval and the evolution of its fairness through multiple evaluation cycles. However, we note that methodologies in this area are still emerging.

7.3 Opportunities in Feedback Loops

Feedback can also be harnessed to adjust system performance towards greater fairness. Sonboli *et al.* (2020a) present an adaptive recommendation approach to multidimensional fairness using probabilistic social choice to control subgroup fairness over time. In this model, deviations from fairness observed in a particular time window are addressed by adjusting the system's fairness objectives over the next batch of recommendations produced.

Biega *et al.* (2018) also account for time in their reranking strategy; while their algorithm does not directly use relevance feedback, it considers past rankings so that the ranking at time t improves the aggregate fairness the system achieved up to time t.

8

Next Steps for Fair Information Access

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

But while general literature on algorithmic fairness and fair machine learning is a crucial starting point, information access systems present particular problems and possibilities that make the straightforward application of existing concepts insufficient, as we have shown in Section 4.1. In particular, the multisided nature and ranked outputs of many information access systems complicate the problem of assessing their fairness, as we must identify which stakeholders we are concerned with treating fairly and develop a definition of fairness that applies to repeated, ranked outputs, among other challenges. The work of fair information access often requires data that is not commonly included

with recommender systems or information retrieval data sets, particularly when seeking to ensure results are fair with respect to sensitive characteristics users or creators, such as their gender or ethnicity. This work must also be done with great care and compassion to ensure that users and creators are treated with respect and dignity and to avoid various traps that result in overbroad or ungeneralizable claims.

We argue that there is nothing particularly new about these requirements, but that thinking about the fairness of information access brings to the surface issues that should be considered in all research and development.

8.1 Directions for Future Research

There are many open problems that need attention in fair information access. Some of the ones we see include:

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

- *Deeper study of the development and evolution of biases over time.* Most work — with the exception of fair policy learning and a handful of other studies — focuses on one-shot batch evaluation of information access systems and their fairness. However, system behavior is dynamic over time as the system processes information requests, produces results, users respond to them, and the system

learns from their feedback. This dynamicism means that an initially fair system may become unfair over time if users respond to it in a biased or discriminatory fashion, or that it may move towards a more fair state if users respond well to recommendations that increase overall fairness. Tools such as T-RECS Lucherini *et al.* (2021) may be valuable for such research.

- *Define and study further fairness concerns beyond consumer and provider fairness.* We have identified subject fairness as one additional type of concern here, but we doubt it is the only additional stakeholder whose equity concerns should be considered.

- *Study human desires for and response to fairness interventions in information access.* The first works are beginning to surface in this direction (Smith *et al.*, 2020), Harambam *et al.* (2019) explored users' desired features and capabilities for recommendation with concerns that touch on fairness, and Ferraro *et al.* (2021) studied provider perceptions, but at present little is known about what users or content providers expect from a system with respect to its fairness, or how users will respond to fairness-enhancing interventions in information access systems.

- *Develop appropriate metrics for information access fairness, along with thorough understanding of the requirements and behavior of fairness metrics and best practices for applying them in practical situations.* For example, we believe expected exposure (Diaz *et al.*, 2020) and pairwise fairness (Beutel *et al.*, 2019) are useful frameworks for reasoning about many provider fairness concerns, but there is a much work left to do to understand how best to apply and interpret them in offline and online studies.

- *Develop standards and best practices for information access data and model provenance.* Gebru *et al.* (2018) presented the idea of *datasheets* for data sets, arguing that data sets should be thoroughly and carefully documented so downstream users can properly assess their applicability, limitations, and the appropriateness of a proposed use. Information retrieval has a long history

of careful attention to evaluation data through TREC, CLEF, and similar initiatives (Voorhees, 2001), but until recently the evaluations in question have typically focused on overall effectiveness with some explorations of related issues such as diversity. Recommender systems has a substantial library of data sets, but has seen less attention to their careful documentation; Harper and Konstan (2015) provide a notable exception in their documentation of the MovieLens data set, addressing in advance several of the questions proposed by Gebru *et al.* (2018).

Mitchell *et al.* (2019) built on this idea for reporting important properties of trained models, and Yang *et al.* (2018) present a "nutrition label" for (non-personalized) rankings describing their data sources, ranking principles, and other information. These concepts need to be extended to information access, and to the complex integrated data sets that drive many search and recommendation applications. New research continues to discover that long-standing data management decisions, such as pruning (Beel and Brunel, 2019), may have deep implications for experiment and recommendation outcomes, emphasizing the need for careful study of the properties of recommendation data, models, and outputs that should be documented.

- *Engage more deeply with the multidimensional and complex nature of bias.* Most of the existing literature on fair information access — and indeed all of algorithmic fairness — focuses on single attributes in isolation, often restricting them to binary values. However, the intersection of group memberships often gives rise to particular forms of discrimination and social bias that cannot be explained by any one of the groups alone (Crenshaw, 1989). Some recent work begins to engage with multiple simultaneous axes of discrimination or fairness (Yang *et al.*, 2020), but as with many mathematical formulations of social concepts, multidimensionality does not fully capture the dynamics invoked by the concept of intersectionality (Hoffmann, 2019). Further, many social categories are complex, unstable, and socially constructed, and algorithmic fairness is only just beginning to reckon with these complexities of human

social experience. Hanna *et al.* (2020) present a treatment of some of these issues in the context of algorithmic fairness, but much work remains to respond to that call and make fairness — both generally for machine learning and specifically for information access — responsive to these realities.

- *Participatory design and research in information access.* Participatory design (Schuler and Namioka, 1993) has a long history in human-computer interaction and user-centered design, but it is difficult to find examples of it applied to the design, evaluation, and study of modern, large-scale information access systems. Belkin and Robertson (1976) observe that "it is necessary to establish and maintain an effective social relationship between [information] science and those whom it affects, so that the latter have a means of judging the implications of the former's activities"; this is true in general, but particularly for the concerns of this monograph. The field is accumulating many techniques for measuring and providing different kinds of fairness, but a serious understanding about what affected people actually want is currently wanting. One notable exception is the work of Harambam *et al.* (2019), who studied what Dutch news consumers want in terms of the control their news recommendation service provides. Smith *et al.* (2020) and Sonboli *et al.* (2021) studied users' opinions of fairness in recommendation, but similar studies of producers, subjects, and other affected stakeholders are needed.

There is a lot of open space for research in fair information access, and this work has the potential for significant improvements to the human and societal impact of algorithmic systems for locating, retrieving, filtering, and ranking information.

8.2 Recommendations for Studying Fairness

Finally, we wish to leave our readers with some suggestions for how to approach research, study, and practice in fair information access, based on the work and concepts we have synthesized in this monograph.

Define the goal. Effective work on fair information access begins with a clear social goal: what specific fairness-related harms are to be avoided? What is the legal, ethical, or other basis for understanding and defining those harms? We hope our map of the space in Section 4 helps in that definitional work. This is crucial for many reasons, but one is to avoid abstraction traps (Selbst *et al.*, 2019) by keeping the work grounded in specific applications and risks; effective and appropriate generalization, in our opinion, flows from clear, contextualized findings.

Clearly operationalize the goal. With a specific harm in mind, select a metric that plausibly captures the kind of harm to be avoided. Project writeups, whether as formal research papers or internal reports, need to clearly and specifically describe how (un)fairness and its resulting harms are being measured, and justify why it is an appropriate means of measuring the target concept. The work we have cited in Sections 5 and 6 provides examples of doing this for various fairness objectives. Jacobs and Wallach (2021) provide a more thorough treatment of the complexities of measuring subjective, contestible constructs like fairness.

Use appropriate data. Data is one of the major challenges for fairness research, in part because group fairness work often requires sensitive data that is often not collected with normal information retrieval or recommender systems data sets. Some data sets provide group annotations, such as the data from the TREC Fair Ranking tracks (Biega *et al.*, 2020) and certain older MovieLens data sets (Harper and Konstan, 2015). For some content creators, library data can be a source of author demographic information (Ekstrand and Kluver, 2021). As noted in Section 1.8.2, we advise against statistical inference techniques for annotating individual people; there has been work, however, on using background distributions to estimate metrics (Kallus *et al.*, 2020).

Carefully report limitations. Any research study has limitations, and fairness studies are no exception. It is crucial to carefully and thoughtfully report the limitations of the data, metrics, and methods in order to help readers appropriately interpret and generalize the results. Bracing

honesty in discussing what any particular work can and cannot do is key to making true progress in this space.

8.3 Concluding Remarks

As we said at the outset (Section 1.6), it is our hope that this monograph provides readers with an information access background who wish to learn about algorithmic fairness, and people grounded in algorithmic fairness and curious about what is happening on fairness in information retrieval, with a good starting point to understand the complexities, pitfalls, and possibilities in the rich and high-impact problem space of fair information access. This field is still young; far too young to provide a comprehensive, retrospective treatment of its key ideas and findings.

What we have sought to do instead is to collect the work so far and integrate it into a prospective map of the space. Much of this map is still incomplete, and the next years of research will fill in many details and likely unlock entirely new dimensions to consider. We look forward to seeing the field grow and reading the many papers to come, and remember, please cite who we cite, not just us.

Acknowledgements

Michael Ekstrand's contributions are based upon work supported by the National Science Foundation under Grant No. IIS 17-51278. Anubrata Das is supported in part by the Micron Foundation, Wipro, and Good Systems[1], a University of Texas at Austin Grand Challenge to develop responsible AI technologies. Robin Burke's work was supported by the National Science Foundation under Grant No. IIS 19-11025. We also thank James Atwood, Asia Biega, Yoni Halpern, Matt Lease, Hansa Srinivasan, and the anonymous reviewers for providing additional feedback and suggestions.

[1] https://goodsystems.utexas.edu/

Appendix

A

Resources for Fair Information Access

In this appendix, we collect pointers to several resources for studying and working on fair information access. We have made every effort to ensure these links are current as of the time of publication, but they may degrade more quickly than the references in the rest of the publication.

A.1 Data Sets

- The TREC Fair Ranking track (launched in 2019) provides data sets for provider fairness in search rankings, both in academic search (2019–2020) and Wikipedia article search (2021). The data is available in TREC (https://trec.nist.gov/results.html), with the track web site at https://fair-trec.github.io.
- The PIReT Book Data Tools at https://bookdata.piret.info provide tools to integrate book recommendation data sets (including from BookCrossing, Amazon, and GoodReads) with publicly-available book and author metadata to study provider fairness in book recommendation, as used by Ekstrand and Kluver (2021).
- Ghosh *et al.* (2021) develop a number of data sets for fair ranking, using various methods and studying the errors of demographic inference for data augmentation.

A.2 Software

There are not yet widely-distributed open-source software for fair recommendation and retrieval; the available code is mostly embedded in published experiment scripts, or general-purpose systems repurposed for fair information access.

- Terrier (http://terrierteam.dcs.gla.ac.uk/research.html) provides xQuAD, a diversification technique that has been successfully applied for fair search ranking (Mcdonald and Ounis, 2020).
- Experimental scripts are available for the fair recommendation studies of Ekstrand and Kluver (2021) (https://md.ekstrandom.net/pubs/bag-extended) and Ekstrand *et al.* (2018b) (https://md.ekstrandom.net/pubs/cool-kids).
- librec-auto (https://librec-auto.readthedocs.io/en/latest/) provides automated support for running recommender systems experiments, including fairness metrics.

References

Abdollahpouri, H. (2020). "Popularity Bias in Recommendation: A Multi-stakeholder Perspective". *PhD thesis*. University of Colorado Boulder. URL: https://arxiv.org/pdf/2008.08551.pdf.

Abdollahpouri, H., G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. (2020). "Multi-stakeholder Recommendation: Survey and Research Directions". en. *User Modeling and User-Adapted Interaction*. 30(1): 127–158. DOI: 10.1007/s11257-019-09256-1.

ACM Council. (2018). "ACM Code of Ethics and Professional Conduct". *Tech. rep.* Association for Computing Machinery. URL: https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct.

Albright, A. (2019). "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions". *Harvard John M. Olin Fellow's Discussion Paper*. 85. URL: https://thelittledataset.com/about_files/albright_judge_score.pdf.

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". *Proceedings of the ACM on Human-Computer Interaction*. 3(CSCW): 1–30. DOI: 10.1145/3359301.

Alstyne, M. van and E. Brynjolfsson. (2005). "Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities". *Management Science*. 51(6): 851–868. DOI: 10.1287/mnsc.1050.0363.

Angwin, J., J. Larson, L. Kirchner, and S. Mattu. (2016). "Machine Bias". URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aziz, H. (2019). "Developments in Multi-Agent Fair Allocation". co.RR Nov. arXiv: 1911.09852 [cs.GT].

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. CIKM '08*. New York, NY, USA: Association for Computing Machinery. 561–570.

Barocas, S. and A. D. Selbst. (2016). "Big Data's Disparate Impact". *California Law Review*. 104(3): 671. DOI: 10.15779/Z38BG31.

Baeza-Yates, R. (2018). "Bias on the web". en. *Communications of the ACM*. May. DOI: 10.1145/3209581.

Bagdasaryan, E., O. Poursaeed, and V. Shmatikov. (2019). "Differential Privacy Has Disparate Impact on Model Accuracy". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc. 15479–15488. URL: http://papers.nips.cc/paper/9681-differential-privacy-has-disparate-impact-on-model-accuracy.pdf.

Barocas, S., A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, W. D. Wadsworth, and H. Wallach. (2021). "Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21*. Virtual Event, USA: Association for Computing Machinery. 368–378. DOI: 10.1145/3461702.3462610.

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

Basilico, J. and Y. Raimond. (2017). "Déjà Vu: The Importance of Time and Causality in Recommender Systems". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems. RecSys '17*. Como, Italy: Association for Computing Machinery. 342. DOI: 10.1145/3109859.3109922.

Becker, C. D. and E. Ostrom. (1995). "HUMAN ECOLOGY AND RESOURCE SUSTAINABILITY: The Importance of Institutional Diversity". *Annual Review of Ecology and Systematics*. 26(1): 113–133. DOI: 10.1146/annurev.es.26.110195.000553.

Becker, H. (1982). *Art Worlds*. University of California Press.

Beel, J. and V. Brunel. (2019). "Data Pruning in Recommender Systems Research: Best-Practice or Malpractice?" In: *ACM RecSys 2019 Late-Breaking Results*. URL: http://ceur-ws.org/Vol-2431/paper6.pdf.

Belkin, N. J. and S. E. Robertson. (1976). "Some ethical and political implications of theoretical research in information science". In: *Proceedings of the ASIS Annual Meeting*.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21*. Virtual Event, Canada: Association for Computing Machinery. 610–623. DOI: 10.1145/3442188.3445922.

Beutel, A., J. Chen, Z. Zhao, and E. H. Chi. (2017). "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations". July. arXiv: 1707.00075 [cs.LG].

Beutel, A., E. H. Chi, C. Goodrow, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, and L. Hong. (2019). "Fairness in Recommendation Ranking through Pairwise Comparisons". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM Press. DOI: 10.1145/3292500.3330745.

Biega, A. J., F. Diaz, M. D. Ekstrand, and S. Kohlmeier. (2020). "Overview of the TREC 2019 Fair Ranking Track". In: *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*. URL: https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.FR.pdf.

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*. ACM. 405–414. DOI: 10.1145/3209978.3210063.

Bigdeli, A., N. Arabzadeh, S. Seyedsalehi, M. Zihayat, and E. Bagheri. (2021). "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. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1748–1752. DOI: 10.1145/3404835.3463110.

Binns, R. (2020). "On the Apparent Conflict between Individual and Group Fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 514–524. DOI: 10.1145/3351095.3372864.

Bodenhausen, G. V. and M. Lichtenstein. (1987). "Social Stereotypes and Information-Processing Strategies: The Impact of Task Complexity". *Journal of Personality and Social Psychology*. 52(5): 871–880. DOI: 10.1037/0022-3514.52.5.871.

Bolukbasi, T., K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. Ed. by D. D. Lee and M. Sugiyama and U. V. Luxburg and I. Guyon and R. Garnett. Curran Associates, Inc. URL: http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.

Browne, S. (2015). *Dark Matters: On the Surveillance of Blackness*. en. Duke University Press. URL: https://play.google.com/store/books/details?id=snmJCgAAQBAJ.

Budish, E. and E. Cantillon. (2012). "The Multi-unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard". *The American Economic Review*. 102(5): 2237–2271.

Buolamwini, J. and T. Gebru. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*. Vol. 81. *Proceedings of Machine Learning Research*. PMLR. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

Burke, R. (2017). "Multisided Fairness for Recommendation". July. arXiv: 1707.00093 [cs.CY].

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*. Ed. by S. A. Friedler and C. Wilson. Vol. 81. *Proceedings of Machine Learning Research*. New York, NY, USA: PMLR. 202–214. URL: http://proceedings.mlr.press/v81/burke18a.html.

Burke, V. I. and R. D. Burke. (2019). "Powerlessness and Personalization". *The International Journal of Applied Philosophy*. 33(2): 319–343. DOI: 10.5840/ijap202034131.

Campos, P. G., F. Díez, and I. Cantador. (2014). "Time-Aware Recommender Systems: A Comprehensive Survey and Analysis of Existing Evaluation Protocols". *User Modeling and User-Adapted Interaction*. 24(1): 67–119. DOI: 10.1007/s11257-012-9136-x.

Cañamares, R. and P. Castells. (2018). "Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: ACM. 415–424. DOI: 10.1145/3209978.3210014.

Carbonell, J. and J. Goldstein. (1998). "The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98*. New York, NY, USA: ACM. 335–336. DOI: 10.1145/290941.291025.

Carterette, B. (2011). "System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11*. Beijing, China: Association for Computing Machinery. 903–912. DOI: 10.1145/2009916.2010037.

Caton, S. and C. Haas. (2020). "Fairness in Machine Learning: A Survey". Oct. arXiv: 2010.04053 [cs.LG].

Celis, L. E. and V. Keswani. (2019). "Improved Adversarial Learning for Fair Classification". Jan. arXiv: 1901.10443 [cs.LG].

Celis, L. E. and V. Keswani. (2020). "Implicit Diversity in Image Summarization". *Proceedings of the ACM on Human-Computer Interaction*. 4(CSCW2): 1–28. DOI: 10.1145/3415210.

Celis, L. E., D. Straszak, and N. K. Vishnoi. (2018). "Ranking with Fairness Constraints". In: *45th International Colloquium on Automata, Languages, and Programming*. Ed. by I. Chatzigiannakis, C. Kaklamanis, D. Marx, and D. Sannella. Vol. 107. *Leibniz International Proceedings in Informatics (LIPIcs)*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany. DOI: 10.4230/LIPIcs.ICALP.2018.28.

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. NETFLIX '08*. No. Article 5. Las Vegas, Nevada: Association for Computing Machinery. 1–8. DOI: 10.1145/1722149.1722154.

Chakraborty, A., A. Hannak, A. Biega, and K. Gummadi. (2017). "Fair Sharing for Sharing Economy Platforms". *Fairness, Accountability and Transparency in Recommender Systems*. Aug. URL: http://scholarworks.boisestate.edu/fatrec/2017/1/6.

Chandar, P., F. Diaz, and B. St. Thomas. (2020). "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems". URL: https://github.com/pchandar/beyond-accuracy-tutorial.

Chaney, A. J. B., 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. RecSys '18*. Vancouver, British Columbia, Canada: Association for Computing Machinery. 224–232. DOI: 10.1145/3240323.3240370.

Chen, I., F. D. Johansson, and D. Sontag. (2018). "Why Is My Classifier Discriminatory?" In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc. 3539–3550. URL: http://papers.nips.cc/paper/7613-why-is-my-classifier-discriminatory.pdf.

Cho, J., S. Roy, and R. E. Adams. (2005). "Page Quality: In Search of an Unbiased Web Ranking". In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. SIGMOD '05*. Baltimore, Maryland: Association for Computing Machinery. 551–562. DOI: 10.1145/1066157.1066220.

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. DOI: 10.1086/669608.

Chouldechova, A. (2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". en. *Big Data*. 5(2): 153–163. arXiv: 1610.07524 [stat.AP]. URL: http://dx.doi.org/10.1089/big.2016.0047.

Crawford, K. (2017). "The Trouble with Bias". Neural Information Processing Systems 2017. URL: https://youtu.be/fMym_BKWQzk.

Crenshaw, K. (1989). "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". *The University of Chicago legal forum*. 1989: 139–168. URL: https://heinonline.org/HOL/P?h=hein.journals/uchclf1989&i=143.

Croft, W. B., D. Metzler, and T. Strohman. (2010). *Search Engines: Information Retrieval in Practice*. Pearson Education, Inc.

Cui, S., W. Pan, C. Zhang, and F. Wang. (2021). "Towards Model-Agnostic Post-Hoc Adjustment for Balancing Ranking Fairness and Algorithm Utility". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21*. Virtual Event, Singapore: Association for Computing Machinery. 207–217. DOI: 10.1145/3447548.3467251.

D'Amour, A., H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern. (2020). "Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 525–534. DOI: 10.1145/3351095.3372878.

D'Ignazio, C. and L. F. Klein. (2020). *Data Feminism*. MIT Press. URL: https://data-feminism.mitpress.mit.edu/.

Das, A. and M. Lease. (2019). "A Conceptual Framework for Evaluating Fairness in Search". July. arXiv: 1907.09328 [cs.IR].

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. (1990). "Indexing by Latent Semantic Analysis". *Journal of the American Society for Information Science*. 41(6): 391–407. URL: http://www3.interscience.wiley.com/journal/10049585/abstract?CRETRY=1&SRETRY=0.

Deldjoo, Y., V. W. Anelli, H. Zamani, A. Bellogin, and T. Di Noia. (2019). "Recommender Systems Fairness Evaluation via Generalized Cross Entropy". In: *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments at RecSys '19*. Vol. 2440. CEUR-WS. URL: http://arxiv.org/abs/1908.06708.

Deshpande, M. and G. Karypis. (2004). "Item-based Top-N Recommendation Algorithms". *ACM Transactions on Information Systems*. 22(1): 143–177. DOI: 10.1145/963770.963776.

Diakopoulos, N. (2015). "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures". en. *Digital Journalism*. 3(3): 398–415. DOI: 10.1080/21670811.2014.976411.

Diaz, F., B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. (2020). "Evaluating Stochastic Rankings with Expected Exposure". In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. ACM. DOI: 10.1145/3340531.3411962.

Dragovic, N., I. Madrazo Azpiazu, and M. S. Pera. (2016). ""Is Sven Seven?": A Search Intent Module for Children". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16*. Pisa, Italy: ACM. 885–888. DOI: 10.1145/2911451.2914738.

Dutta, S., D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. (2020). "Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing". 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. 2803–2813. URL: http://proceedings.mlr.press/v119/dutta20a.html.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. (2012). "Fairness Through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ITCS '12*. Cambridge, Massachusetts: Association for Computing Machinery. 214–226. DOI: 10.1145/2090236.2090255.

Dwork, C. and C. Ilvento. (2018). "Fairness Under Composition". In: *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Ed. by A. Blum. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. DOI: 10.4230/LIPICS.ITCS.2019.33.

Edwards, H. and A. Storkey. (2016). "Censoring Representations with an Adversary". In: *Proceedings of the International Conference on Learning Representations (ICLR 2016)*. URL: http://arxiv.org/abs/1511.05897.

Ekstrand, M. D., R. Joshaghani, and H. Mehrpouyan. (2018a). "Privacy for All: Ensuring Fair and Equitable Privacy Protections". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by S. A. Friedler and C. Wilson. Vol. 81. *Proceedings of Machine Learning Research*. New York, NY: PMLR. 35–47. URL: https://proceedings.mlr.press/v81/ekstrand18a.html.

Ekstrand, M. D. and D. Kluver. (2021). "Exploring Author Gender in Book Rating and Recommendation". *User Modeling and User-Adapted Interaction*. 31(3). DOI: 10.1007/s11257-020-09284-2.

Ekstrand, M. D., M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. (2018b). "All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by S. A. Friedler and C. Wilson. Vol. 81. *Proceedings of Machine Learning Research*. New York, New York: PMLR. 172–186. URL: https://proceedings.mlr.press/v81/ekstrand18b.html.

Ekstrand, M. D., M. Tian, M. R. I. Kazi, H. Mehrpouyan, and D. Kluver. (2018c). "Exploring Author Gender in Book Rating and Recommendation". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. Vancouver British Columbia Canada: ACM. DOI: 10.1145/3240323.3240373.

Ensign, D., S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. (2018). "Runaway Feedback Loops in Predictive Policing". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by S. A. Friedler and C. Wilson. Vol. 81. *Proceedings of Machine Learning Research*. New York, NY, USA: PMLR. 160–171. URL: http://proceedings.mlr.press/v81/ensign18a.html.

Epps-Darling, A., R. T. Bouyer, and H. Cramer. (2020). "Artist Gender Representation in Music Streaming". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR. 248–254. URL: https://program.ismir2020.net/poster_2-11.html.

Epstein, R. and R. E. Robertson. (2015). "The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections". en. *Proceedings of the National Academy of Sciences of the United States of America*. 112(33): E4512–21. DOI: 10.1073/pnas.1419828112.

Evans, D., R. Schmalensee, M. Noel, H. Chang, and D. Garcia-Swartz. (2011). *Platform Economics: Essays on Multi-Sided Businesses*. Competition Policy International. URL: http://ssrn.com/abstract=1974020.

Evans, D. S. and R. Schmalensee. (2016). *Matchmakers: The New Economics of Multisided Platforms*. Harvard Business Review Press.

Feldman, M., S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. (2015). "Certifying and Removing Disparate Impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 259–268. DOI: 10.1145/2783258.2783311.

Feng, R., Y. Yang, Y. Lyu, C. Tan, Y. Sun, and C. Wang. (2019). "Learning Fair Representations via an Adversarial Framework". Apr. arXiv: 1904.13341 [cs.LG].

Ferraro, A. (2019). "Music Cold-Start and Long-Tail Recommendation: Bias in Deep Representations". In: *Proceedings of the 13th ACM Conference on Recommender Systems. RecSys '19*. Copenhagen, Denmark: Association for Computing Machinery. 586–590. DOI: 10.1145/3298689.3347052.

Ferraro, A., X. Serra, and C. Bauer. (2021). "What is Fair? Exploring the Artists' Perspective on the Fairness of Music Streaming Platforms". In: *Proceedings of the 18th IFIP International Conference on Human-Computer Interaction (INTERACT 2021)*. URL: http://arxiv.org/abs/2106.02415.

Ferro, N., N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindn, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. (2018). "The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction". *SIGIR Forum*. 52(1): 91–101. DOI: 10.1145/3274784.3274789.

Fetterly, D. and Z. Gyöngyi, eds. (2009). *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*. Madrid, Spain: Association for Computing Machinery. URL: https://dl.acm.org/doi/proceedings/10.1145/1531914.

Finn, M. and Q. DuPont. (2020). "From closed world discourse to digital utopianism: the changing face of responsible computing at Computer Professionals for Social Responsibility (1981–1992)". *Internet Histories*. 4(1): 6–31.

Fisher, E. (2022). "Do algorithms have a right to the city? Waze and algorithmic spatiality". en. *Cultural Studies of Science Education*. 36(1): 74–95. DOI: 10.1080/09502386.2020.1755711.

Fleder, D. M. 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. DOI: 10.1287/mnsc.1080.0974.

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)*. 1918–1921. DOI: 10.1109/ICDE48307.2020.00203.

Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian. (2016). "On the (im)possibility of fairness". Sept. arXiv: 1609.07236 [cs.CY].

Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian. (2021). "The (Im)possibility of Fairness". en. *Communications of the ACM*. 64(4): 136–143. DOI: 10.1145/3433949.

Friedler, S. A., C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. (2019). "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. Atlanta, GA, USA: Association for Computing Machinery. 329–338. DOI: 10.1145/3287560.3287589.

Friedman, A., B. P. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky. (2015). "Privacy Aspects of Recommender Systems". In: *Recommender Systems Handbook*. Boston, MA: Springer US. 649–688. DOI: 10.1007/978-1-4899-7637-6_19.

Friedman, B. and H. Nissenbaum. (1996). "Bias in Computer Systems". *ACM Transactions on Information Systems*. 14(3): 330–347. DOI: 10.1145/230538.230561.

Funk, S. (2006). "Netflix Update: Try This at Home". URL: http://sifter.org/~simon/journal/20061211.html.

García-Soriano, D. and F. Bonchi. (2021). "Maxmin-Fair Ranking: Individual Fairness under Group-Fairness Constraints". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21*. Virtual Event, Singapore: Association for Computing Machinery. 436–446. DOI: 10.1145/3447548.3467349.

Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. (2018). "Datasheets for Datasets". Mar. arXiv: 1803.09010 [cs.DB].

Geyik, S. C., S. Ambler, and K. Kenthapadi. (2019). "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19*. Anchorage, AK, USA: Association for Computing Machinery. 2221–2231. DOI: 10.1145/3292500.3330691.

Geyik, S. C. and K. Kenthapadi. (2018). "Building Representative Talent Search at LinkedIn". URL: https://engineering.linkedin.com/blog/2018/10/building-representative-talent-search-at-linkedin.

Ghosh, A., R. Dutt, and C. Wilson. (2021). "When Fair Ranking Meets Uncertain Inference". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1033–1043. DOI: 10.1145/3404835.3462850.

Glowacka, D. (2019). "Bandit Algorithms in Information Retrieval". *Foundations and Trends® in Information Retrieval*. 13(4): 299–424. DOI: 10.1561/1500000067.

Goecks, J., A. Voida, S. Voida, and E. D. Mynatt. (2008). "Charitable Technologies: Opportunities for Collaborative Computing in Nonprofit Fundraising". In: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work. CSCW '08*. New York, NY, USA: ACM. 689–698. DOI: 10.1145/1460563.1460669.

Goffman, W. (1964). "A Searching Procedure for Information Retrieval". *Information Storage and Retrieval*. 2(2): 73–78. DOI: 10.1016/0020-0271(64)90006-3.

Goldman, E. (2005). "Search Engine Bias and the Demise of Search Engine Utopianism". *Yale Journal of Law & Technology*. 8: 188+. URL: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=&v=2.1&it=r&id=GALE%7CA182079911&sid=googleScholar&linkaccess=fulltext&userGroupName=anon%7Ee0e23234.

Gómez, E., C. Shui Zhang, L. Boratto, M. Salamó, and M. Marras. (2021). "The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1808–1812. DOI: 10.1145/3404835.3463235.

Gosepath, S. (2011). "Equality". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2011. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2011/entries/equality/.

Green, B. and Y. Chen. (2019). "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. Atlanta, GA, USA: ACM. 90–99. DOI: 10.1145/3287560.3287563.

Guo, R., X. Zhao, A. Henderson, L. Hong, and H. Liu. (2020). "Debiasing Grid-Based Product Search in E-Commerce". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '20*. New York, NY, USA: Association for Computing Machinery. 2852–2860.

Hamidi, F., M. K. Scheuerman, and S. M. Branham. (2018). "Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18*. ACM. 8. DOI: 10.1145/3173574.3173582.

Hammer, M. C. (2021). "You bore us.. . . " URL: https://twitter.com/MCHammer/status/1363908982289559553.

Hanna, A., E. Denton, A. Smart, and J. Smith-Loud. (2020). "Towards a Critical Race Methodology in Algorithmic Fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 501–512. DOI: 10.1145/3351095.3372826.

Harambam, J., D. Bountouridis, M. Makhortykh, and J. van Hoboken. (2019). ". Designing forthe Better by Taking UsersintoAccount:A Qualitative Evaluation of User Control Mechanismsin (News) Recommender Systems". In: *Proceedings of the 13th ACM Conference on Recommender Systems. RecSys '19*. Copenhagen, Denmark: Association for Computing Machinery. 69–77. DOI: 10.1145/3298689.3347014.

Hardt, M., E. Price, and N. Srebro. (2016). "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. papers.nips.cc. 3315–3323. URL: http://papers.nips.cc/paper/6373-equality-of-opportunity-in-supervised-learning.

Harman, J. L., J. O'Donovan, T. Abdelzaher, and C. Gonzalez. (2014). "Dynamics of Human Trust in Recommender Systems". In: *Proceedings of the 8th ACM Conference on Recommender Systems. RecSys '14*. Foster City, Silicon Valley, California, USA: Association for Computing Machinery. 305–308. DOI: 10.1145/2645710.2645761.

Harper, F. M. and J. A. Konstan. (2015). "The MovieLens Datasets: History and Context". *ACM Transactions on Interactive Intelligent Systems*. 5(4): 19:1–19:19. DOI: 10.1145/2827872.

Hashimoto, T., M. Srivastava, H. Namkoong, and P. Liang. (2018). "Fairness Without Demographics in Repeated Loss Minimization". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. *Proceedings of Machine Learning Research*. Stockholmsmässan, Stockholm Sweden: PMLR. 1929–1938. URL: http://proceedings.mlr.press/v80/hashimoto18a.html.

Herlocker, J., J. A. Konstan, and J. Riedl. (2002). "An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms". en. *Information Retrieval*. 5(4): 287–310. DOI: 10.1023/A:1020443909834.

Hoffmann, A. L. (2019). "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse". *Information, Communication and Society*. 22(7): 900–915. DOI: 10.1080/1369118X.2019.1573912.

HUD. (2020). "HUD's Implementation of the Fair Housing Act's Disparate Impact Standard". *Federal Register*. 85(186): 60288–60333. URL: https://www.federalregister.gov/documents/2020/09/24/2020-19887/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard.

Hurley, N. and M. Zhang. (2011). "Novelty and Diversity in Top-N Recommendation – Analysis and Evaluation". *ACM Transactions on Internet Technology*. 10(4): 14:1–14:30. DOI: 10.1145/1944339.1944341.

Hutchinson, B. and M. Mitchell. (2019). "50 Years of Test (Un)fairness: Lessons for Machine Learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. Atlanta, GA, USA: Association for Computing Machinery. 49–58. DOI: 10.1145/3287560.3287600.

Hutchinson, B., A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. (2021). "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21*. Virtual Event, Canada: Association for Computing Machinery. 560–575. DOI: 10.1145/3442188.3445918.

IFLA Governing Board. (2012). "Code of Ethics for Librarians and Other Information Workers". *Tech. rep.* International Federation of Library Associations and Institutions. URL: https://repository.ifla.org/handle/123456789/1850.

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. FAccT '21*. Virtual Event, Canada: Association for Computing Machinery. 375–385. DOI: 10.1145/3442188.3445901.

Jambor, T., J. Wang, and N. Lathia. (2012). "Using Control Theory for Stable and Efficient Recommender Systems". In: *Proceedings of the 21st International Conference on World Wide Web. WWW '12*. 11–20. DOI: 10.1145/2187836.2187839.

Jannach, D., L. Lerche, I. Kamehkhosh, and M. Jugovac. (2015). "What Recommenders Recommend: An Analysis of Recommendation Biases and Possible Countermeasures". en. *User Modeling and User-Adapted Interaction*. 25(5): 427–491. DOI: 10.1007/s11257-015-9165-3.

Jiang, R., S. Chiappa, T. Lattimore, A. György, and P. Kohli. (2019). "Degenerate feedback loops in recommender systems". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay. (2017). "Accurately Interpreting Clickthrough Data as Implicit Feedback". *SIGIR Forum*. 51(1): 4–11. DOI: 10.1145/3130332.3130334.

Johnson, I., J. Henderson, C. Perry, J. Schöning, and B. Hecht. (2017). "Beautiful... but at What Cost? An Examination of Externalities in Geographic Vehicle Routing". *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 1(2): 1–21. DOI: 10.1145/3090080.

Jones, R. and K. L. Klinkner. (2008). "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs". In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08*. Napa Valley, California, USA: Association for Computing Machinery. 699–708. DOI: 10.1145/1458082.1458176.

Joseph, M., M. Kearns, J. Morgenstern, S. Neel, and A. Roth. (2018). "Meritocratic Fairness for Infinite and Contextual Bandits". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18*. New Orleans, LA, USA: Association for Computing Machinery. 158–163. DOI: 10.1145/3278721.3278764.

Joseph, M., M. Kearns, J. H. Morgenstern, and A. Roth. (2016). "Fairness in Learning: Classic and Contextual Bandits". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc. 325–333. URL: http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf.

Kallus, N., X. Mao, and A. Zhou. (2020). "Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 110. DOI: 10.1145/3351095.3373154.

Kamiran, F. and T. Calders. (2009). "Classifying Without Discriminating". In: *2009 2nd International Conference on Computer, Control and Communication*. ieeexplore.ieee.org. 1–6. DOI: 10.1109/IC4.2009.4909197.

Kamiran, F. and T. Calders. (2012). "Data Preprocessing Techniques for Classification Without Discrimination". en. *Knowledge and Information Systems*. 33(1): 1–33. DOI: 10.1007/s10115-011-0463-8.

Kamishima, T. and S. Akaho. (2017). "Considerations on Recommendation Independence for a Find-Good-Items Task". In: *Workshop on Fairness, Accountability and Transparency in Recommender Systems at RecSys 2017*. URL: http://scholarworks.boisestate.edu/fatrec/2017/1/11/.

Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma. (2018). "Recommendation Independence". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by S. A. Friedler and C. Wilson. Vol. 81. *Proceedings of Machine Learning Research*. New York, NY, USA: PMLR. 187–201. URL: http://proceedings.mlr.press/v81/kamishima18a.html.

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. UMAP '18*. Singapore, Singapore: Association for Computing Machinery. 23–28. DOI: 10.1145/3213586.3226206.

Karimi, M., D. Jannach, and M. Jugovac. (2018). "News Recommender Systems — Survey and Roads Ahead". *Information Processing & Management*. 54(6): 1203–1227. DOI: 10.1016/j.ipm.2018.04.008.

Katell, M., M. Young, D. Dailey, B. Herman, V. Guetler, A. Tam, C. Bintz, D. Raz, and P. M. Krafft. (2020). "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 45–55. DOI: 10.1145/3351095.3372874.

Kay, J., B. Kummerfeld, and P. Lauder. (2002). "Personis: A Server for User Models". en. In: *Adaptive Hypermedia and Adaptive Web-Based Systems*. Ed. by P. D. Bra, P. Brusilovsky, and R. Conejo. *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 203–212. URL: http://link.springer.com/chapter/10.1007/3-540-47952-X_22.

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. CHI '15*. Seoul, Republic of Korea: ACM. 3819–3828. DOI: 10.1145/2702123.2702520.

Kaya, M., D. Bridge, and N. Tintarev. (2020). "Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. Virtual Event, Brazil: Association for Computing Machinery. 101–110. DOI: 10.1145/3383313.3412232.

Kearns, M., S. Neel, A. Roth, and Z. S. Wu. (2017). "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness". *arXiv preprint arXiv:1711. 05144*.

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. FAT* '19*. Atlanta, GA, USA: Association for Computing Machinery. 100–109. DOI: 10.1145/3287560.3287592.

Kelly, D. and J. Teevan. (2003). "Implicit Feedback for Inferring User Preference: A Bibliography". *SIGIR Forum*. 37(2): 18–28. DOI: 10.1145/959258.959260.

Khenissi, S., B. Mariem, and O. Nasraoui. (2020). "Theoretical Modeling of the Iterative Properties of User Discovery in a Collaborative Filtering Recommender System". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. Virtual Event, Brazil: Association for Computing Machinery. 348–357. DOI: 10.1145/3383313.3412260.

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. WWW '21*. Ljubljana, Slovenia: Association for Computing Machinery. 1065–1075. DOI: 10.1145/3442381.3450080.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. (2018). "Human Decisions and Machine Predictions". *The Quarterly Journal of Economics*. 133(1): 237–293.

Kleinberg, J., S. Mullainathan, and M. Raghavan. (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Ed. by C. H. Papadimitriou. Vol. 67. *Leibniz International Proceedings in Informatics (LIPIcs)*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany. DOI: 10.4230/LIPICS.ITCS.2017.43.

Kohavi, R., D. Tang, and Y. Xu. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press. DOI: 10.1017/9781108653985.

Kohler-Hausmann, I. (2019). "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination". *Northwestern Law Review*. 113(5): 1163–1228. DOI: 10.2139/ssrn.3050650.

Koren, Y., R. Bell, and C. Volinsky. (2009). "Matrix Factorization Techniques for Recommender Systems". *Computer*. 42(8): 30–37. DOI: 10.1109/MC.2009.263.

Koren, Y. (2010). "Collaborative filtering with temporal dynamics". *Communications of the ACM*. 53(4): 89–97. DOI: 10.1145/1721654.1721677.

Kouki, P., I. Fountalis, N. Vasiloglou, X. Cui, E. Liberty, and K. Al Jadda. (2020). "From the Lab to Production: A Case Study of Session-Based Recommendations in the Home-Improvement Domain". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. Virtual Event Brazil: ACM. 140–149. DOI: 10.1145/3383313.3412235.

Kuhlman, C., W. Gerych, and E. Rundensteiner. (2021). "Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21*. Virtual Event, USA: Association for Computing Machinery. 674–682. DOI: 10.1145/3461702.3462588.

Kuhlthau, C. C. (1993). "A Principle of Uncertainty for Information Seeking". *Journal of Documentation*. 49(4): 339–355. DOI: 10.1108/eb026918.

Kulshrestha, J., M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. (2017). "Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. CSCW '17*. Portland, Oregon, USA: Association for Computing Machinery. 417–432. DOI: 10.1145/2998181.2998321.

Kulynych, B., R. Overdorf, C. Troncoso, and S. Gürses. (2020). "POTs: Protective Optimization Technologies". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 177–188. DOI: 10.1145/3351095.3372853.

Lahoti, P., K. P. Gummadi, and G. Weikum. (2019). "iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 1334–1345. DOI: 10.1109/ICDE.2019.00121.

Lathia, N., S. Hailes, and L. Capra. (2009). "Evaluating Collaborative Filtering Over Time". In: *SIGIR '09 Workshop on the Future of IR Evaluation*. URL: http://www.cs.ucl.ac.uk/staff/n.lathia/papers/lathia_ireval09.pdf.

Lee, M. K., A. Jain, H. J. Cha, S. Ojha, and D. Kusbit. (2019). "Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation". *Proceedings of the ACM on Human-Computer Interaction*. 3(CSCW): 1–26. DOI: 10.1145/3359284.

Lerner, M. (2021). "Fannie Mae to include rent payments in mortgage applicants' credit history review". *Washington Post*. URL: https://www.washingtonpost.com/business/2021/09/08/fannie-mae-include-rent-payments-mortgage-applicants-credit-history-review/.

Leuski, A. and J. Allan. (2000). "Lighthouse: Showing the Way to Relevant Information". In: *Proceedings of the IEEE Symposium on Information Visualization*. 125–130.

Li, L., W. Chu, J. Langford, and R. E. Schapire. (2010). "A Contextual-Bandit Approach to Personalized News Article Recommendation". In: *Proceedings of the 19th International Conference on World Wide Web*. 661–670.

Li, Y., H. Chen, S. Xu, Y. Ge, and Y. Zhang. (2021). "Towards Personalized Fairness based on Causal Notion". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1054–1063. DOI: 10.1145/3404835.3462966.

Lian, J. W., N. Mattei, R. Noble, and T. Walsh. (2018). "The Conference Paper Assignment Problem: Using Order Weighted Averages to Assign Indivisible Goods". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11484.

Lipton, Z., J. McAuley, and A. Chouldechova. (2018). "Does Mitigating ML's Impact Disparity Require Treatment Disparity?" In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc. 8125–8135. URL: http://papers.nips.cc/paper/8035-does-mitigating-mls-impact-disparity-require-treatment-disparity.pdf.

Liu, J., C. Liu, and N. J. Belkin. (2020). "Personalization in Text Information Retrieval: A Survey". *Journal of the Association for Information Science and Technology*. 71(3): 349–369. DOI: 10.1002/asi.24234.

Liu, T.-Y. (2007). "Learning to Rank for Information Retrieval". en. *Foundations and Trends® in Information Retrieval*. 3(3): 225–331. DOI: 10.1561/1500000016.

Longo, B. (2015). *Edmund Berkeley and the Social Responsibility of Computer Professionals*. Association for Computing Machinery and Morgan & Claypool.

Lucherini, E., M. Sun, A. Winecoff, and A. Narayanan. (2021). "T-RECS: A simulation tool to study the societal impact of recommender systems". July. arXiv: 2107.08959 [cs.CY].

Luhn, H. P. (1960). "Key Word-in-Context Index for Technical Literature (KWIC Index)". *American Documentation*. 11(4): 288–295. DOI: 10.1002/asi.5090110403.

Madras, D., E. Creager, T. Pitassi, and R. Zemel. (2018). "Learning Adversarially Fair and Transferable Representations". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. *Proceedings of Machine Learning Research*. PMLR. 3384–3393. URL: https://proceedings.mlr.press/v80/madras18a.html.

Mann, T. A., S. Gowal, A. Gyorgy, H. Hu, R. Jiang, B. Lakshminarayanan, and P. Srinivasan. (2019). "Learning from Delayed Outcomes via Proxies with Applications to Recommender Systems". 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. 4324–4332.

Manning, C. D., P. Raghavan, and H. Schütze. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mansoury, M., H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, and R. Burke. (2020). "Feedback Loop and Bias Amplification in Recommender Systems". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Virtual Event Ireland: ACM. 2145–2148. DOI: 10.1145/3340531.3412152.

Mcdonald, G. and I. Ounis. (2020). "University of Glasgow Terrier Team at the TREC 2020 Fair Ranking Track". In: *The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings*. Vol. 1266. *NIST SP*. URL: https://trec.nist.gov/pubs/trec29/papers/uogTr.FR.pdf.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. (2019). "A survey on bias and fairness in machine learning. arXiv". *arXiv preprint arXiv:1908. 09635*.

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. WWW '17 Companion*. Perth, Australia: International World Wide Web Conferences Steering Committee. 626–633. DOI: 10.1145/3041021.3054197.

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: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. Torino, Italy: Association for Computing Machinery. 2243–2251. DOI: 10.1145/3269206.3272027.

Metaxa, D., M. A. Gan, S. Goh, J. Hancock, and J. A. Landay. (2021). "An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations". *Proceedings of the ACM on Human-Computer Interaction*. 5(CSCW1): 1–23. DOI: 10.1145/3449100.

Metzler, D. and T. Kanungo. (2008). "Machine Learned Sentence Selection Strategies for Query-Biased Summarization". In: *SIGIR Learning to Rank Workshop*.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum. (2020). "Algorithmic Fairness: Choices, Assumptions, and Definitions". en. *Annual Review of Statistics and Its Application*. 8(Nov.). DOI: 10.1146/annurev-statistics-042720-125902.

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. (2019). "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. Atlanta, GA, USA: Association for Computing Machinery. 220–229. DOI: 10.1145/3287560.3287596.

Moffat, A. and J. Zobel. (2008). "Rank-biased Precision for Measurement of Retrieval Effectiveness". *ACM Transactions on Information Systems*. 27(1): 2:1–2:27. DOI: 10.1145/1416950.1416952.

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*. New York, NY, USA: Association for Computing Machinery. 429–438. DOI: 10.1145/3397271.3401100.

Moulin, H. (2004). *Fair Division and Collective Welfare*. en. MIT Press. URL: https://play.google.com/store/books/details?id=qQXtEnb2B2cC.

Mowshowitz, A. and A. Kawaguchi. (2002). "Assessing Bias in Search Engines". en. *Information Processing & Management*. 38(1): 141–156. DOI: 10.1016/s0306-4573(01)00020-6.

Nagel, R. (2021). "Went to Google (first mistake) to see..." URL: https://twitter.com/rebeccanagle/status/1371535405942734849.

Narasimhan, H., A. Cotter, M. Gupta, and S. Wang. (2020). "Pairwise Fairness for Ranking and Regression". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 5248–5255. DOI: 10.1609/aaai.v34i04.5970.

Nasr, M. and M. C. Tschantz. (2020). "Bidding Strategies with Gender Nondiscrimination Constraints for Online Ad Auctions". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 337–347. DOI: 10.1145/3351095.3375783.

Nielek, R., A. Wierzbicki, A. Jatowt, and K. Tanaka. (2016). "Report on the WebQuality 2015 Workshop". In: *ACM SIGIR Forum*. Vol. 50. 83–85.

Ning, X. and G. Karypis. (2011). "SLIM: Sparse Linear Methods for Top-N Recommender Systems". In: *Proceedings of the 2011 IEEE 11th International Conference on Data Mining. ICDM '11*. Washington, DC, USA: IEEE Computer Society. 497–506. DOI: 10.1109/ICDM.2011.134.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. en. NYU Press. URL: https://market.android.com/details?id=book--ThDDwAAQBAJ.

Noriega-Campero, A., M. A. Bakker, B. Garcia-Bulle, and A. '. Pentland. (2019). "Active Fairness in Algorithmic Decision Making". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES '19*. Honolulu, HI, USA: Association for Computing Machinery. 77–83. DOI: 10.1145/3306618.3314277.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. en. Penguin Books. URL: http://www.worldcat.org/oclc/1015602855.

Okonofua, J. A. and J. L. Eberhardt. (2015). "Two strikes: Race and the disciplining of young students". *Psychological science*. 26(5): 617–624.

Okonofua, J. A., G. M. Walton, and J. L. Eberhardt. (2016). "A Vicious Cycle: A Social-Psychological Account of Extreme Racial Disparities in School Discipline". *Perspectives on psychological science: a journal of the Association for Psychological Science*. 11(3): 381–398.

Olteanu, A., C. Castillo, F. Diaz, and E. Kıcıman. (2019). "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". en. *Frontiers in Big Data*. 2(July): 13. DOI: 10.3389/fdata.2019.00013.

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. CHI '17*. Denver, Colorado, USA: Association for Computing Machinery. 6620–6631. DOI: 10.1145/3025453.3025727.

Pandey, S. and C. Olston. (2008). "Crawl Ordering by Search Impact". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining. WSDM '08*. Palo Alto, California, USA: Association for Computing Machinery. 3–14. DOI: 10.1145/1341531.1341535.

Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. en. Penguin.

Pasquale, F. (2006). "Rankings, Reductionism, and Responsibility". *Cleveland State Law Review*. URL: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/clevslr54&section=10&casa_token=BlLTyPvQ06wAAAAA:8Xa_ROPfsnjS9x52QtvwocH1adCMbyENgLbwKrnWyscAtulNRpk9oKwAaegjm0oWDxiFFcVhr-0.

Patro, G. K., A. Chakraborty, A. Banerjee, and N. Ganguly. (2020). "Towards Safety and Sustainability: Designing Local Recommendations for Post-pandemic World". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. Virtual Event, Brazil: Association for Computing Machinery. 358–367. DOI: 10.1145/3383313.3412251.

Pelly, L. (2018). "Discover Weakly". *The Baffler*. June.

Pizzato, L., T. Rej, T. Chung, I. Koprinska, and J. Kay. (2010). "RECON: a reciprocal recommender for online dating". In: *Proceedings of the fourth ACM conference on Recommender systems. RecSys '10*. Barcelona, Spain: Association for Computing Machinery. 207–214. DOI: 10.1145/1864708.1864747.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. (2017). "On Fairness and Calibration". In: *Proceedings of the 31st Conference on Neural Information Processing Systems*. URL: https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf.

Raghavan, M., S. Barocas, J. Kleinberg, and K. Levy. (2020). "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 469–481. DOI: 10.1145/3351095.3372828.

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*. ACM Press. DOI: 10.1145/3477495.3532018.

Raj, A., A. Milton, and M. D. Ekstrand. (2021). "Pink for Princesses, Blue for Superheroes: The Need to Examine Gender Stereotypes in Kid's Products in Search and Recommendations". In: *Proceedings of the 5th International and Interdisciplinary Perspectives on Children and Recommender and Information Retrieval Systems: Search and Recommendation Technology through the Lens of a Teacher (KidRec 2021), co-located with ACM IDC 2021*. URL: http://arxiv.org/abs/2105.09296.

Raj, A., C. Wood, A. Montoly, and M. D. Ekstrand. (2020). "Comparing Fair Ranking Metrics". Sept. arXiv: 2009.01311 [cs.IR].

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*. Virtual Event, Canada: Association for Computing Machinery. 306–316. DOI: 10.1145/3404835.3462949.

Rendle, S., C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. (2009). "BPR: Bayesian Personalized Ranking from Implicit Feedback". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. UAI '09*. Arlington, Virginia, United States: AUAI Press. 452–461. URL: http://dl.acm.org/citation.cfm?id=1795114.1795167.

Rich, E. (1979). "User Modeling via Stereotypes". *Cognitive Science*. 3(4): 329–354. URL: http://www.sciencedirect.com/science/article/B6W48-4FWF9GC-9/2/f924f793eb153d455893e8d39982ef45.

Rijsbergen, C. J. van. (1979). *Information Retrieval*. Butterworths.

Robertson, S. E. (1977). "The Probability Ranking Principle in IR". *Journal of Documentation*. 33(4): 294–304. DOI: 10.1108/eb026647.

Rochet, J.-C. and J. Tirole. (2003). "Platform Competition in Two-Sided Markets". *Journal of the European Economic Association*. 1(4): 990–1029.

Roegiest, A., A. Lipani, A. Beutel, A. Olteanu, A. Lucic, A.-A. Stoica, A. Das, A. Biega, B. Voorn, C. Hauff, D. Spina, D. Lewis, D. W. Oard, E. Yilmaz, F. Hasibi, G. Kazai, G. McDonald, H. Haned, I. Ounis, I. van der Linden, J. Garcia-Gathright, J. Baan, K. N. Lau, K. Balog, M. de Rijke, M. Sayed, M. Panteli, M. Sanderson, M. Lease, M. D. Ekstrand, P. Lahoti, and T. Kamishima. (2019). "FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval". *SIGIR Forum*. 53(2): 20–43. URL: http://sigir.org/wp-content/uploads/2019/december/p020.pdf.

Rolf, E., T. T. Worledge, B. Recht, and M. Jordan. (2021). "Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. *Proceedings of Machine Learning Research*. PMLR. 9040–9051. URL: https://proceedings.mlr.press/v139/rolf21a.html.

Rossetti, M., F. Stella, and M. Zanker. (2016). "Contrasting Offline and Online Results when Evaluating Recommendation Algorithms". In: *Proceedings of the 10th ACM Conference on Recommender Systems. RecSys '16*. New York, NY, USA: ACM. 31–34. DOI: 10.1145/2959100.2959176.

Roth, A. E. (2015). *Who Gets What and Why: The New Economics of Matchmaking and Market Design*. Houghton Mifflin Harcourt.

Rothstein, R. (2017). *The Color of Law: A Forgotten History of How Our Government Segregated America*. en. Liveright Publishing. URL: https://play.google.com/store/books/details?id=SdtDDQAAQBAJ.

Salimi, B., L. Rodriguez, B. Howe, and D. Suciu. (2019). "Interventional Fairness: Causal Database Repair for Algorithmic Fairness". In: *Proceedings of the 2019 International Conference on Management of Data. SIGMOD '19*. Amsterdam, Netherlands: Association for Computing Machinery. 793–810. DOI: 10.1145/3299869.3319901.

Salton, G., A. Wong, and C. S. Yang. (1975). "A Vector Space Model for Automatic Indexing". *Communications of the ACM*. 18(11): 613–620. DOI: 10.1145/361219.361220.

Sambasivan, N., E. Arnesen, B. Hutchinson, and V. Prabhakaran. (2020). "Non-portability of Algorithmic Fairness in India". Dec. arXiv: 2012.03659 [cs.CY].

Sánchez-Monedero, J., L. Dencik, and L. Edwards. (2020). "What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery. 458–468. DOI: 10.1145/3351095.3372849.

Santos, R. L. T., C. Macdonald, and I. Ounis. (2015). *Search Result Diversification*. Vol. 9. *Foundations and Trends in Information Retrieval*. Hanover, MA, USA: Now Publishers Inc.

Santos, R. L. T., J. Peng, C. Macdonald, and I. Ounis. (2010). "Explicit Search Result Diversification through Sub-queries". In: *ECIR 2010: Advances in Information Retrieval*. Vol. 5993. *LNCS*. Springer Berlin Heidelberg. 87–99. DOI: 10.1007/978-3-642-12275-0_11.

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. DOI: 10.1145/3308560.3317595.

Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock. (2002). "Methods and Metrics for Cold-start Recommendations". In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '02*. Tampere, Finland: ACM. 253–260. DOI: 10.1145/564376.564421.

Schuler, D. and A. Namioka, eds. (1993). *Participatory Design: Principles and Practices*. English. Hillsdale, N.J: CRC / Lawrence Erlbaum Associates. URL: http://www.amazon.com/Participatory-Design-Principles-Douglas-Schuler/dp/0805809511?ie=UTF8&keywords=participatory%20design&qid=1460985157&ref_=sr_1_2&sr=8-2.

Selbst, A. D., D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. (2019). "Fairness and Abstraction in Sociotechnical Systems". en. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. Atlanta, GA, USA: Association for Computing Machinery. 59–68. DOI: 10.1145/3287560.3287598.

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. KDD '18*. London, United Kingdom: ACM. 2219–2228. DOI: 10.1145/3219819.3220088.

Singh, A. and T. Joachims. (2019). "Policy Learning for Fairness in Ranking". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc. 5426–5436. URL: http://papers.nips.cc/paper/8782-policy-learning-for-fairness-in-ranking.pdf.

Singh, V. K., M. Chayko, R. Inamdar, and D. Floegel. (2020). "Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms". en. *Journal of the Association for Information Science and Technology*. 71(11): 1281–1294. DOI: 10.1002/asi.24335.

Singhal, A., C. Buckley, and M. Mitra. (2017). "Pivoted Document Length Normalization". *SIGIR Forum*. 51(2): 176–184. DOI: 10.1145/3130348.3130365.

Smith, J., N. Sonbolil, C. Fiesler, and R. Burke. (2020). "Exploring User Opinions of Fairness in Recommender Systems". In: *Fair & Responsible AI Workshop @ CHI 2020*. URL: https://fair-ai.owlstown.net/publications/1459.

Smucker, M. D. and J. Allan. (2006). "Find-Similar: Similarity Browsing as a Search Tool". 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. 461–468. DOI: 10.1145/1148170.1148250.

Sonboli, N. (2022). "Controlling the Fairness / Accuracy Tradeoff in Recommender Systems". *PhD thesis*. University of Colorado.

Sonboli, N., R. Burke, N. Mattei, F. Eskandanian, and T. Gao. (2020a). ""And the Winner Is...": Dynamic Lotteries for Multi-group Fairness-Aware Recommendation". arXiv: 2009.02590 [cs.IR].

Sonboli, N., F. Eskandanian, R. Burke, W. Liu, and B. Mobasher. (2020b). "Opportunistic Multi-aspect Fairness through Personalized Re-ranking". In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20*. Genoa, Italy: Association for Computing Machinery. 239–247. DOI: 10.1145/3340631.3394846.

Sonboli, N., J. J. Smith, F. Cabral Berenfus, R. Burke, and C. Fiesler. (2021). "Fairness and Transparency in Recommendation: The Users' Perspective". In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '21*. Utrecht, Netherlands: Association for Computing Machinery. 274–279. DOI: 10.1145/3450613.3456835.

Speer, R. (2017). "ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors". URL: https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/.

Srivastava, M., H. Heidari, and A. Krause. (2019). "Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19*. Anchorage, AK, USA: Association for Computing Machinery. 2459–2468. DOI: 10.1145/3292500.3330664.

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

Steenkamp, J.-B. E. M. and H. Baumgartner. (1998). "Assessing Measurement Invariance in Cross-National Consumer Research". en. *The Journal of consumer research*. 25(1): 78–90. DOI: 10.1086/209528.

Stray, J. (2020). "Aligning AI Optimization to Community Well-Being". *International Journal of Community Well-Being*. 3(4): 443–463. DOI: 10.1007/s42413-020-00086-3.

Sun, W., O. Nasraoui, and P. Shafto. (2018). "Iterated Algorithmic Bias in the Interactive Machine Learning Process of Information Filtering". In: *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Seville, Spain: SCITEPRESS - Science and Technology Publications. 108–116. DOI: 10.5220/0006938301100118.

Sun, Z., D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, and C. Geng. (2020). "Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. Virtual Event, Brazil: Association for Computing Machinery. 23–32. DOI: 10.1145/3383313.3412489.

Suresh, H. and J. V. Guttag. (2019). "A Framework for Understanding Unintended Consequences of Machine Learning". Jan. arXiv: 1901.10002 [cs.LG].

Takács, G. and D. Tikk. (2012). "Alternating Least Squares for Personalized Ranking". In: *Proceedings of the Sixth ACM conference on Recommender Systems. RecSys '12*. Dublin, Ireland: Association for Computing Machinery. 83–90. DOI: 10.1145/2365952.2365972.

Thelwall, M. (2019). "Reader and Author Gender and Genre in GoodReads". *Journal of Librarianship and Information Science*. 51(2): 403–430. DOI: 10.1177/0961000617709061.

Thomson, W. (2016). "Introduction to the Theory of Fair Allocation". In: *Handbook of Computational Social Choice*. Ed. by F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. Cambridge University Press. 261–283. DOI: 10.1017/CBO9781107446984.012.

Thune, J. (2019). "S.2763 Filter Bubble Transparency Act". URL: https://www.congress.gov/bill/116th-congress/senate-bill/2763.

Tintarev, N. and J. Masthoff. (2007). "A Survey of Explanations in Recommender Systems". In: *IEEE ICDE Workshop*. IEEE Computer Society. 801–810. DOI: 10.1109/ICDEW.2007.4401070.

Tombros, A. and M. Sanderson. (1998). "Advantages of Query Biased Summaries in Information Retrieval". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98*. Melbourne, Australia: Association for Computing Machinery. 2–10. DOI: 10.1145/290941.290947.

Vaughan, L. and M. Thelwall. (2004). "Search Engine Coverage Bias: Evidence and Possible Causes". *Information Processing & Management*. 40(4): 693–707. DOI: 10.1016/S0306-4573(03)00063-3.

Vaughan, L. and Y. Zhang. (2007). "Equal Representation by Search Engines? A Comparison of Websites across Countries and Domains". *Journal of computer-mediated communication: JCMC*. 12(3): 888–909. DOI: 10.1111/j.1083-6101.2007.00355.x.

Voorhees, E. M. (2001). "The Philosophy of Information Retrieval Evaluation". en. In: *Evaluation of Cross-Language Information Retrieval Systems*. Ed. by C. Peters, M. Braschler, J. Gonzalo, and M. Kluck. *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 355–370. URL: http://link.springer.com/chapter/10.1007/3-540-45691-0_34.

Wu, Y., J. Cao, G. Xu, and Y. Tan. (2021). "TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1013–1022. DOI: 10.1145/3404835.3462882.

Wyden, R. (2019). "S.1108 Algorithmic Accountability Act of 2019". URL: https://www.congress.gov/bill/116th-congress/senate-bill/1108.

Xiang, A. and I. D. Raji. (2019). "On the Legal Compatibility of Fairness Definitions". Nov. arXiv: 1912.00761 [cs.CY].

Xie, X., J. Mao, Y. Liu, M. de Rijke, Y. Shao, Z. Ye, M. Zhang, and S. Ma. (2019). "Grid-Based Evaluation Metrics for Web Image Search". In: *The World Wide Web Conference. WWW '19*. San Francisco, CA, USA: Association for Computing Machinery. 2103–2114. DOI: 10.1145/3308558.3313514.

Xu, D., Y. Wu, S. Yuan, L. Zhang, and X. Wu. (2019). "Achieving Causal Fairness Through Generative Adversarial Networks". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Yang, K., J. R. Loftus, and J. Stoyanovich. (2020). "Causal Intersectionality for Fair Ranking". June. arXiv: 2006.08688 [cs.LG].

Yang, K. and J. Stoyanovich. (2017). "Measuring Fairness in Ranked Outputs". In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. No. Article 22. Chicago, IL, USA: ACM. 1–6. DOI: 10.1145/3085504.3085526.

Yang, K., J. Stoyanovich, A. Asudeh, B. Howe, H. V. Jagadish, and G. Miklau. (2018). "A Nutritional Label for Rankings". en. In: *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18*. Houston, TX, USA: ACM Press. 1773–1776. DOI: 10.1145/3183713.3193568.

Yao, S. and B. Huang. (2017). "Beyond Parity: Fairness Objectives for Collaborative Filtering". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc. 2925–2934. URL: http://papers.nips.cc/paper/6885-beyond-parity-fairness-objectives-for-collaborative-filtering.pdf.

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. DOI: 10.1145/1772690.1772793.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. (2017). "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web. WWW '17*. Perth, Australia: International World Wide Web Conferences Steering Committee. 1171–1180. DOI: 10.1145/3038912.3052660.

Zehlike, M., F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. (2017). "FA*IR: A Fair Top-k Ranking Algorithm". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17*. ACM. 1569–1578. DOI: 10.1145/3132847.3132938.

Zehlike, M., K. Yang, and J. Stoyanovich. (2022). "Fairness in Ranking, Part I: Score-based Ranking". *ACM Computing Surveys*. Apr. DOI: 10.1145/3533379.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. (2013). "Learning Fair Representations". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. *Proceedings of Machine Learning Research*. Atlanta, Georgia, USA: PMLR. 325–333. URL: https://proceedings.mlr.press/v28/zemel13.html.

Zhang, J., G. Adomavicius, A. Gupta, and W. Ketter. (2020). "Consumption and Performance: Understanding Longitudinal Dynamics of Recommender Systems via an Agent-Based Simulation Framework". *Information Systems Research*. 31(1): 76–101. DOI: 10.1287/isre.2019.0876.

Zhang, X., M. Khaliligarekani, C. Tekin, and M. Liu. (2019). "Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc. 15269–15278. URL: http://papers.nips.cc/paper/9662-group-retention-when-using-machine-learning-in-sequential-decision-making-the-interplay-between-user-dynamics-and-fairness.pdf.

Zhao, X., Z. Niu, and W. Chen. (2013). "Opinion-Based Collaborative Filtering to Solve Popularity Bias in Recommender Systems". In: *Database and Expert Systems Applications*. Springer Berlin Heidelberg. 426–433. DOI: 10.1007/978-3-642-40173-2_35.

Zheng, G., F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li. (2018). "DRN: A Deep Reinforcement Learning Framework for News Recommendation". In: *Proceedings of the 2018 World Wide Web Conference. WWW '18*. Lyon, France: International World Wide Web Conferences Steering Committee. 167–176. DOI: 10.1145/3178876.3185994.

Zhu, Z., J. Kim, T. Nguyen, A. Fenton, and J. Caverlee. (2021). "Fairness among New Items in Cold Start Recommender Systems". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 767–776. DOI: 10.1145/3404835.3462948.

Ziegler, C.-N., S. M. McNee, J. A. Konstan, and G. Lausen. (2005). "Improving Recommendation Lists Through Topic Diversification". In: *Proceedings of the 14th international conference on World Wide Web. WWW '05*. Chiba, Japan: Association for Computing Machinery. 22–32. DOI: 10.1145/1060745.1060754.

Index