Fairness in Ranking, Part I: Score-Based Ranking

MEIKE ZEHLIKE, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany
KE YANG, New York University, NY, and University of Massachusetts, Amherst, MA, USA
JULIA STOYANOVICH, New York University, NY, USA

In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers, with contributions coming from the data management, algorithms, information retrieval, and recommender systems communities. In this survey, we give a systematic overview of this work, offering a broad perspective that connects formalizations and algorithmic approaches across sub-fields. An important contribution of our work is in developing a common narrative around the value frameworks that motivate specific fairness-enhancing interventions in ranking. This allows us to unify the presentation of mitigation objectives and of algorithmic techniques to help meet those objectives or identify trade-offs.

In this first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this article, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In the second part of this survey, we present methods that incorporate fairness in supervised learning, and also give representative examples of recent work on fairness in recommendation and matchmaking systems. We also discuss evaluation frameworks for fair score-based ranking and fair learning-to-rank, and draw a set of recommendations for the evaluation of fair ranking methods.

CCS Concepts: • **Information systems** → **Data management systems**; • **Social and professional topics** → **Computing/technology policy;**

Additional Key Words and Phrases: Fairness, ranking, set selection, responsible data science, survey

ACM Reference format:
Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.* 55, 6, Article 118 (December 2022), 36 pages.
https://doi.org/10.1145/3533379

1 INTRODUCTION

The research community recognizes several important normative dimensions of information technology including privacy, transparency, and fairness. In this survey, we focus on fairness—a broad and inherently interdisciplinary topic of which the social and philosophical foundations are still unresolved [17].

This research was supported in part by NSF Awards No. 1934464, 1916505, and 1922658.
Authors' addresses: M. Zehlike, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany; email: meikezehlike@mpi-sws.org; K. Yang, New York University, NY, and University of Massachusetts, Amherst, MA, USA; email: ky630@nyu.edu; J. Stoyanovich, New York University, NY, USA; email: stoyanovich@nyu.edu.

0360-0300/2022/12-ART118 $15.00
https://doi.org/10.1145/3533379

Research on fair machine learning has mainly focused on classification and prediction tasks [8, 17], while we focus on ranking. As is customary in fairness research, we assume that input data describes *individuals*—natural persons seeking education, employment, or financial opportunities, or being prioritized for access to goods and services. While some of the algorithmic techniques described here can be applied to entities other than people, we believe that the concept of fairness, along with the corresponding normative frameworks, applies pre-dominantly to scenarios where data describes people. For consistency, we will refer to the set of individuals in the input to a ranking task as *candidates*.

We consider two types of ranking tasks: score-based and supervised learning. In score-based ranking, a given set of candidates is sorted on the score attribute, which may itself be computed on the fly, and returned in sorted order. In supervised learning-to-rank, a preference-enriched training set of candidates is given, with preferences among them stated in the form of scores, preference pairs, or lists; this training set is used to train a model that predicts the ranking of unseen candidates. For both score-based and supervised learning tasks, we typically return the best-ranked k candidates, the top-k. Set selection is a special case of ranking that ignores the relative order among the top-k, returning them as a set.

While supervised learning-to-rank appears to be similar to classification, there is one crucial difference. The goal of classification is to assign a class label to each item, and this assignment is made independently for each item. In contrast, learning-to-rank positions items relative to each other, and so the outcome for one item is not independent of the outcomes for the other items. This lack of independence has profound implications for the design of learning-to-rank methods in general, and for fair learning-to-rank in particular.

To make our discussion concrete, we now present our running example from university admissions, a domain in which ranking and set selection are very natural and are broadly used.

1.1 Running Example: University Admissions

Consider an admissions officer at a university who selects candidates from a large applicant pool. When making their decision, the officer pursues some or all of the goals listed below. Some of these goals may be legally mandated, while others may be based on the policies adopted by the university, and include admitting students who:

- —are likely to succeed: complete the program with high marks and graduate on time;
- —show strong interest in specific majors like computer science, art, or literature; and
- —form a demographically diverse group in terms of their demographics, both overall and in each major.

Figure 1 shows a dataset C of applicants and illustrates the admissions process. Each applicant submits several quantitative scores, all transformed here to a discrete scale of 1 (worst) through 5 (best) for ease of exposition: X_1 is the high school **grade point average** (**GPA**), X_2 is the verbal portion of the **Scholastic Assessment Test**(**SAT**) score, and X_3 is the mathematics portion of the SAT score. Attribute X_4 (choice) is a weighted feature vector extracted from the applicant's essay, with weight ranging between 0 and 1, and with a higher value corresponding to stronger interest in a specific major. For example, candidate b is a White male with a high GPA (4 out of 5), perfect SAT verbal and SAT math scores (5 out of 5), a strong interest in studying computer science (feature weight 0.9), and some interest in studying art (weight 0.2).

The admissions officer uses a suite of tools to sift through the applications and identify promising candidates. Many of these tools are *rankers*, illustrated in Figure 3. A ranker takes a dataset of candidates, described by structured features, text, or both, as input and produces a permutation of these candidates, also called a *ranking*. The admissions officer will take the order in which the

candidate	A_1	A_2	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
b	male	White	4	5	5	{cs:0.9; art:0.2}	14	9	1
c	male	Asian	5	3	4	{math:0.9; cs:0.5}	12	9	1
d	female	White	5	4	2	{lit:0.8; math:0.8}	11	4	6
e	male	White	3	3	4	{math:0.8; econ:0.4}	10	7	6
f	female	Asian	3	2	3	{econ:0.9; math:0.5}	8	5	8
k	female	Black	2	2	3	{lit:0.9;art:0.8}	7	1	9
l	male	Black	1	1	4	{lit:0.5; math:0.7}	6	6	2
o	female	White	1	1	2	{econ:0.9; cs:0.8}	4	7	8

(a)

τ_1
b
c
d
e
f
k
l
o

(b)

τ_2
c
b
e
f
d
o
l
k

(c)

τ_3
k
l
b
d
e
f
c
o

(d)

Fig. 1. (a) Dataset C of college applicants, with demographic attributes A_1 (sex) and A_2 (race), numerical attributes X_1 (high school GPA), X_2 (verbal SAT), and X_3 (math SAT), and attribute X_4 (choice), that is, a vector extracted from the applicants' essays; (b) is a ranking $\boldsymbol{\tau}_1$ on Y_1, computed as the sum of X_1, X_2, and X_3; (c) is a ranking on Y_2, predicted based on historical performance of STEM (cs, econ, math) majors; (d) is a ranking on Y_3, predicted based on historical performance of humanities (art, lit) majors. In all cases, the top-4 candidates will be interviewed in score order, and potentially admitted.

candidates appear in a ranking under advisement when deciding whom to consider more closely, interview, and admit.

These tools include score-based rankers (Section 2.1) that compute the score of each candidate based on a formula that the admissions officer gives, and then return some number of highest-scoring applicants in ranked order. This *scoring formula* may, for example, specify the score as a linear combination of the applicant's high-school GPA and the two components of their SAT score, each carrying an equal weight. This is done in Figure 1(a), where a candidate's score is computed as $Y_1 = X_1 + X_2 + X_3$ and then ranking $\boldsymbol{\tau}_1$ in Figure 1(b) is produced.

Predictive analytics are also among the admissions officer's toolkit. For example, multiple ranking models may be trained, one per undergraduate major or set of majors, on features X_1, X_2, X_3, X_4 of the successful applicants from the past years, to predict applicant's standing upon graduation (based, e.g., on their GPA in the major). These ranking models are then used to predict a ranking of this year's applicants. In our example in Figure 1(a), feature Y_2 predicts performance in a STEM (Science, Technology, Engineering, and Mathematics) major such as computer science (cs), economics (econ), or mathematics (math) and leads to ranking $\boldsymbol{\tau}_2$ in Figure 1(c), while feature Y_3 predicts performance in a humanities major such as literature (lit) or fine arts (art) and leads to ranking $\boldsymbol{\tau}_3$ in Figure 1(d).

The promising applicants identified in this way—with the help of either a score-based ranker or a predictive analytic—will then be considered more closely, *in ranked order*: invited for an interview and potentially admitted.

Let us recall that, in addition to incorporating quantitative scores and students' choice, an admissions officer also aims to admit a demographically diverse group of students to the university and to each major. Furthermore, the admissions officer is increasingly aware that the data on which their decisions are based may be biased, in the sense that this data may carry results of historical discrimination or disadvantage, and that the computational tools at their disposal may be exacerbating or introducing new forms of bias, or even creating a kind of a self-fulfilling prophecy. (See discussion of the types of bias in Section 3.2.) For this reason, the officer may elect to incorporate one or several fairness objectives into the ranking process.

For example, they may assert, for legal or ethical reasons, that the proportion of the female applicants among those selected for further consideration should match their proportion in the input. Applying this requirement to ranking $\boldsymbol{\tau}_1$ in Figure 2 (in which we elaborate on the

candidate	A_1	A_2	X_1	X_2	X_3	Y
b	male	White	4	5	5	14
c	male	Asian	5	3	4	12
d	female	White	5	4	2	11
e	male	White	3	3	4	10
f	female	Asian	3	2	3	8
k	female	Black	2	2	3	7
l	male	Black	1	1	4	6
o	female	White	1	1	2	4

τ_1	τ_2	τ_3
b	b	b
c	c	d
d	d	c
e	f	f
f	e	e
k	k	k
l	l	l
o	o	o

Fig. 2. A dataset C of college applicants. Score Y is computed by a score-based ranker $f(X) = X_1 + X_2 + X_3$. Ranking τ_1 of Y in C. Ranking τ_2 with proportional representation by sex at the top-4. Ranking τ_3 with proportional representation by sex in every prefix of the top-4. The top-4 candidates will be interviewed in score order and potentially admitted.

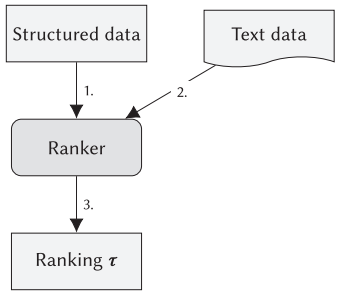

Fig. 3. Functional principle of rankers: (1. and 2.) Structured and text data that correspond to candidates serve as inputs to a ranker; (3.) The ranker outputs a ranking of the candidates τ.

already familiar example in Figure 1) yields ranking τ_2 in Figure 2. Furthermore, the admissions officer may assert that, because applicants are interviewed in ranked order, it is important to achieve proportional representation by sex in *every prefix* of the produced ranking, which yields ranking τ_3 in Figure 2. In this survey, we give an overview of the technical work that would allow an admissions officer to compute ranked results under these and other fairness requirements.

1.2 Scope and Contributions of the Survey

In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers. And while several surveys on fairness in classification have been published (e.g., [48, 54], ranking has not yet received systematic attention. Giving an overview of this large and growing body of work, and the underlying value frameworks that serve as a basis for classification, is the primary goal of our survey. Which specific fairness requirements an admissions officer will assert depends on the values they are operationalizing and, thus, on the mitigation objectives. An important goal of this survey is to create an explicit mapping between mitigation objectives, which we will characterize in Section 3.3. Without such a mapping, an admissions officer in our running example would have a difficult time selecting an appropriate fairness-enhancing intervention, and would not know which interventions are mutually comparable and which are not.

In our survey, we will present a selection of approaches for fairness in ranking that were developed in several sub-fields of computer science, including data management, algorithms, information retrieval, and recommender systems. We are aware of several recent tutorials on fairness in ranking at SIGIR 2019 [14], RecSys 2019 [30], and VLDB 2020 [5], pointing to the need to systematize the work in this area and motivating our survey. Our goal is to offer a broad perspective, connecting work across sub-fields. We discuss existing technical methods for fairness in score-based ranking in Section 5. Technical work on fairness in supervised learning, with a focus on information retrieval, is covered in the second part of the survey, where we also highlight representative examples of fairness in recommender systems and matching.

The primary focus of this survey is on associational fairness measures for ranking, although we do include one recently proposed causal framework.

1.3 Survey Roadmap (Parts I and II)

Part I of this survey is organized as follows:

— We gave a general introduction in Section 1.
— We start with the preliminaries and fix notation in Section 2.
— We present classification frameworks along which we relate all surveyed technical methods in Section 3.
— We present the evaluation datasets that are used by the surveyed technical methods in Section 4.
— We describe technical work on fairness in score-based ranking in Section 5.
— We summarize Part I in Section 6.

Part II of this survey is organized as follows:

— We introduce Part II of the survey in Section 1.
— We recap the relevant notation in Section 2.
— We describe technical work on fair supervised learning in Section 3.
— We highlight representative work on fairness in recommender systems and matching in Section 4.
— We discuss evaluation frameworks for fair score-based ranking and fair learning-to-rank in Section 5.
— We draw a set of recommendations for the evaluation of fair ranking methods in Section 6.
— We conclude the survey, and identify directions for future work, in Section 7.

2 PRELIMINARIES AND NOTATION

In this section, we will build on our running example to discuss score-based and supervised learning-based rankers more formally, and fix the necessary notation. We summarize notation in Table 1 and illustrate it throughout this section.

2.1 Score-based Ranking

Formally, we are given a set C of candidates; each candidate is described by a set of features $\mathbf{X}$ and a score attribute Y. Additionally, we are given a set of sensitive attributes $\mathbf{A} \subseteq \mathbf{X}$, which are categorical, denoting membership of a candidate in demographic groups. Sensitive attributes like age or degree of disability may be drawn from a continuous domain, and several fairness-in-classification methods for continuous sensitive attributes have been proposed [36, 47]. However, we are not aware of any work of this kind that applies to fairness in ranking, and so will assume that sensitive attributes are categorical in the remainder of this survey. A sensitive attribute $A \in \mathbf{A}$ may be binary, with one of the values (e.g., $A = 1$ or $A = \texttt{female}$, as in Figure 1) denoting membership in a minority or historically disadvantaged group (often called "protected group") and with the other value (e.g., $A = 0$ or $A = \texttt{male}$) denoting membership in a majority or privileged group. Alternatively, a sensitive attribute may take on three or more values, for example, to represent ethnicity or (non-binary) gender identity of candidates.

A *ranking* $\boldsymbol{\tau}$ is a permutation over the candidates in C. Letting $n = |C|$, we denote by $\boldsymbol{\tau} = \langle \tau_1, \ldots, \tau_n \rangle$ a ranking that places candidate τ_i at rank i. We denote by $\boldsymbol{\tau}(i)$ the candidate at rank i, and by $\boldsymbol{\tau}^{-1}(a)$ the rank of candidate a in $\boldsymbol{\tau}$. We are often interested in a sub-ranking of $\boldsymbol{\tau}$ containing its best-ranked k elements, for some integer $k \leq n$; this sub-ranking is called the top-k and is denoted $\boldsymbol{\tau}_{1\ldots k}$. For example, given a ranking $\boldsymbol{\tau} = \langle b, c, d, e, f, k, l, o \rangle$, $\boldsymbol{\tau}(3) = d$, $\boldsymbol{\tau}^{-1}(l) = 7$, and the top-4 is $\boldsymbol{\tau}_{1\ldots 4} = \langle b, c, d, e \rangle$.

Table 1. Summary of Notation Used Throughout the Survey

C	A set of candidates to be ranked	Y	the score feature and ground truth for supervised learning
a, b, c	Candidates in C	$\hat{Y}$	the scores predicted by $\hat{f}$
n	Number of candidates $\|C\|$	Y_a	the score of candidate a
$\mathbf{X}$	a set of features of the candidates in C	$\boldsymbol{\tau}$	Ranking: permutation of candidates from C
$\mathbf{X}_a$	Features of candidate a	$\boldsymbol{\tau}(i)$	The candidate at position i in $\boldsymbol{\tau}$
$\mathbf{A}$	A set of sensitive features, $\mathbf{A} \subseteq \mathbf{X}$	$\mathbf{v}(i)$	the position bias of rank i
$\mathcal{G}$	A group (subset) of candidates, $\mathcal{G} \subseteq C$	$U^k(\boldsymbol{\tau})$	Utility of the top-k candidates in $\boldsymbol{\tau}$
$\mathcal{G}_1$	A protected group (subset) of candidates, $\mathcal{G}_1 \subseteq C$	$U^k(\boldsymbol{\tau}, \mathcal{G})$	Utility of the top-k candidates of group $\mathcal{G}$ in $\boldsymbol{\tau}$
$\mathcal{U}$	A set of users that use the ranking system	$U(\boldsymbol{\tau}, a)$	Utility of candidate a in $\boldsymbol{\tau}$
$\mathbf{Q}$	A set of queries	$D(a, b)$	Disparity in visibility between candidates a and b
$f, \hat{f}$	a ranker, a ranker learned from training data	$D(\mathcal{G}_1, \mathcal{G}_2)$	Disparity in visibility between groups $\mathcal{G}_1$ and $\mathcal{G}_2$

Utility. Because score Y is assumed to encode a candidate's appropriateness, quality, or *utility*, a score-based ranking usually satisfies:

$$Y_{\boldsymbol{\tau}(1)} \geq Y_{\boldsymbol{\tau}(2)} \geq \cdots \geq Y_{\boldsymbol{\tau}(n)}. \tag{1}$$

We will find it convenient to denote by $U^k(\boldsymbol{\tau})$ the utility of $\boldsymbol{\tau}_{1\ldots k}$. Different methods surveyed in this article adopt different notions of utility, and we will make their formulations precise as appropriate. The simplest method is to treat $\boldsymbol{\tau}_{1\ldots k}$ as a set (disregarding candidate positions), and to compute the utility of the set as the sum of scores of its elements:

$$U^k(\boldsymbol{\tau}) = \sum_{i=1}^{k} Y_{\boldsymbol{\tau}(i)}. \tag{2}$$

Another common method incorporates position-based discounts, following the observation that it is more important to present high-quality items at the top of the ranked list, since these items are more likely to attract the attention of the consumer of the ranking. For example, we may compute position-discounted utility of a ranking as:

$$U^k(\boldsymbol{\tau}) = \sum_{i=1}^{k} \frac{Y_{\boldsymbol{\tau}(i)}}{\log_2(i+1)}. \tag{3}$$

For example, the utility at top-4 of $\boldsymbol{\tau}_1$ in Figure 2 is 47 based on Equation (2) and 31.4 based on Equation (3). Note that the base of the logarithm in the denominator of Equation (3) is empirically determined, and it can be set to some value $b > 1$ other than 2.

For these variants of utility and for others, it is often useful to quantify utility realized by candidates belonging to a particular demographic group $\mathcal{G} \subseteq C$, defined by an assignment of values to one or several sensitive attributes. For example, $\mathcal{G}$ may contain female candidates, or Asian female candidates. We can then compute to the utility of $\boldsymbol{\tau}_{1\ldots k}$ (per Equation (2)) for group $\mathcal{G}$ as:

$$U^k(\boldsymbol{\tau}, \mathcal{G}) = \sum_{i=1}^{k} Y_{\boldsymbol{\tau}(i)} \times \mathbb{1}[\boldsymbol{\tau}(i) \in \mathcal{G}]. \tag{4}$$

Here, $\mathbb{1}$ is an indicator variable that returns 1 when $\boldsymbol{\tau}(i) \in \mathcal{G}$ and 0 otherwise. Position-discounted utility (per Equation (3)) for group $\mathcal{G}$ can be defined analogously. For the ranking $\boldsymbol{\tau}_1$ in Figure 2, $U^4(\boldsymbol{\tau}_1, sex = \texttt{male}) = 36$, $U^4(\boldsymbol{\tau}_1, sex = \texttt{male} \wedge race = \texttt{White}) = 24$, and $U^4(\boldsymbol{\tau}_1, sex = \texttt{male} \wedge race = \texttt{Black}) = 0$.

Fairness. To satisfy objectives other than utility, such as *fairness*, we may output a ranking $\hat{\boldsymbol{\tau}}$ that is not simply sorted based on the observed values of Y as in Equation (1). As is the case for classification and prediction, numerous fairness measures have been defined for rankings. These

measures can be used both to assess the fairness of a ranking and to intervene on unfairness, for example, by serving as basis for constraints.

A prominent class of fairness measures corresponds to *proportional representation* in the top-k treated as a set, or in every prefix of the top-k. These measures are motivated by the need to mitigate different types of bias, based on assumptions about its origins and with a view of specific objectives (to be discussed in Section 3). For example, ranking $\boldsymbol{\tau}_2$ in Figure 2 re-ranks candidates to satisfy proportional representation by gender at the top-4 (treating it as a set), swapping candidates e and f. The ranking $\boldsymbol{\tau}_3$ in Figure 2 additionally reorders candidates c and d to achieve proportional representation by gender in every prefix of the top-4.

In addition to fairness measures, *diversity* measures have also been proposed in the literature [26]. In this survey, we will discuss coverage-based diversity that is most closely related to fairness, and requires that members of multiple, possibly overlapping, groups, be sufficiently well-represented among the top-k, treated either as a set or as a ranked list. Diversity constraints may, for example, be stated to require that members of each ethnic group, each gender group, and of selected intersectional groups on ethnicity and gender, all be represented at the top-k in proportion to their prevalence in the input. The terminology we adopt in this article is that "coverage-based diversity" is a technical notion that can be used to express several fairness objectives. In contrast, fairness is never purely technical: it is always associated with a value framework and with a socio-technical context of use.

When candidates are re-ranked to meet objectives other than score-based utility, we may be interested to compute *Y-utility loss*, denoted $L_Y(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}})$. We can use a variety of metrics that quantify the distance between ranked lists for this purpose, including, for example, the Kendall distance that counts the number of pairs that appear in the opposite relative order in $\boldsymbol{\tau}$ and $\hat{\boldsymbol{\tau}}$, or one in a family of generalized distances between rankings [42]. However, loss functions that compare rankings $\boldsymbol{\tau}$ and $\hat{\boldsymbol{\tau}}$ in their entirety are uncommon. Rather, Y-utility loss is usually specified over the top-k. The simplest formulation is:

$$L_Y^k(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}}) = U^k(\boldsymbol{\tau}) - U^k(\hat{\boldsymbol{\tau}}). \tag{5}$$

Alternatively, we may normalize this quantity:

$$L_Y^k(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}}) = 1 - \frac{U^k(\hat{\boldsymbol{\tau}})}{U^k(\boldsymbol{\tau})}. \tag{6}$$

Furthermore, we may be interested to quantify utility loss for a particular demographic group $\mathcal{G}$. In that case, we define the utility of $\boldsymbol{\tau}$ and $\hat{\boldsymbol{\tau}}$ for group $\mathcal{G}$, as was done in Equation (4), or analogously for other utility formulations. Interestingly, underrepresented groups may see a gain, rather than a loss, in Y-utility, because they may receive better representation at the top-k when a fairness objective is applied.

3 FOUR CLASSIFICATION FRAMEWORKS FOR FAIRNESS-ENHANCING INTERVENTIONS

Operationally, algorithmic approaches surveyed in this article differ in how they represent candidates (e.g., whether they support one or multiple sensitive attribute, and whether these are binary), in the type of bias they aim to surface and mitigate, in what fairness measure(s) they adopt, in how they navigate the trade-offs between fairness and utility during mitigation, and, for supervised learning methods, at what stage of the pipeline a mitigation is applied. Conceptually, these operational choices correspond to normative statements about the types of bias being observed and mitigated, and the objectives of the mitigation. In this section, we give four classification frameworks that allow us to relate the technical choices with the normative judgments they encode, and to

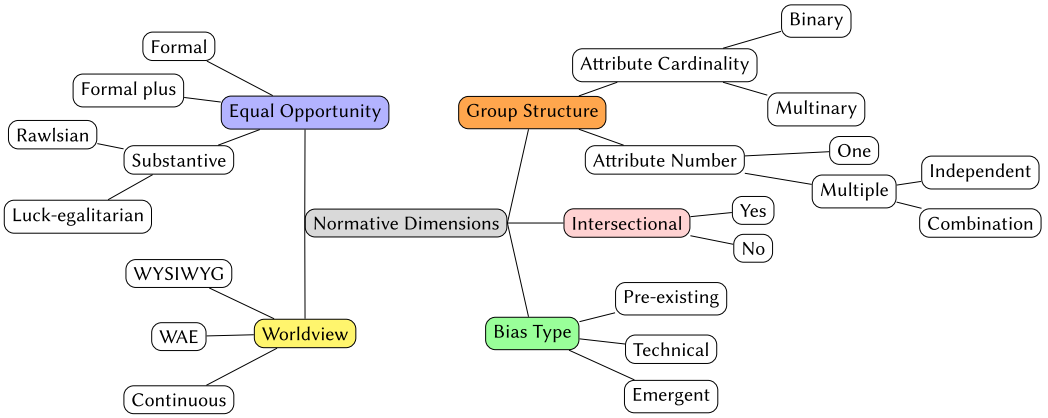

Fig. 4. A mind map summary of the structure of the four classification frameworks.

identify the commonalities and the differences between the many algorithmic approaches. Figure 4 gives a structural overview of the frameworks and their sub-categories in the form of a *mind map*. For each method, we will highlight which normative choices they make based on this mind map.

3.1 Group Structure

Recall that fairness of a method is stated with respect to a set of categorical sensitive attributes (or features). Individuals who have the same value of a particular sensitive attribute, such as gender or race, are called a *group*. In this survey, we consider several orthogonal dimensions of group structure, based on the handling of sensitive attributes.

Cardinality of sensitive attributes. Some methods consider only *binary* sensitive attributes (e.g., binary gender, majority or minority ethnic group), while other methods handle higher-cardinality (*multinary*) domains of values for sensitive attributes. If a multinary domain is supported, methods differ in whether they consider one of the values to be protected (corresponding to a designated group that has been experiencing discrimination), or if they treat all values of the sensitive attribute as potentially being subject to discrimination.

Number of sensitive attributes. Some methods are designed to handle a *single sensitive attribute* at a time (e.g., they handle gender or race, but not both), while other methods handle *multiple sensitive attributes* simultaneously (e.g., they handle both gender and race).

Handling of multiple sensitive attributes. Methods that support multiple sensitive attributes differ in whether they handle these *independently* (e.g., by asserting fairness constraints w.r.t. the treatment of both women and Blacks) or *in combination* (e.g., by requiring fairness w.r.t. Black women). Note that any method that supports a single multinary attribute can be used to represent multiple sensitive attributes with the help of a computed high-cardinality sensitive attribute. For example, a computed sensitive attribute *gender-race-disability* can represent the Cartesian product $\{male, female, non\text{-}binary\} \times \{White, Black, Asian\} \times \{disabled, non\text{-}disabled\}$. We may be tempted to say that such methods take the point of view of intersectional discrimination [23, 46]. However, as we will discuss in Section 3.3, detecting and mitigating intersectional discrimination is more nuanced, and so it is in general not true that if a method takes a Cartesian product of sensitive attribute values then handles intersectional discrimination, and if a method treats sensitive attributes independently then it does not.

3.2 Type of Bias

We study ranking systems with respect to the types of bias that they attempt to mitigate, namely, pre-existing bias, technical bias, and emergent bias, as defined by [35].

Pre-existing bias. This type of bias includes all biases that exist independently of an algorithm itself and has its origins in society. For an example of pre-existing bias in rankings, consider the SAT. College applicants in the US are commonly ranked on their SAT score, often in combination with other features. It has been documented that the mean score of the math section of the SAT differs across racial groups, as does the shape of the score distribution. According to a Brookings report that analyzed 2015 SAT test results, "The mean score on the math section of the SAT for all test-takers is 511 out of 800, the average scores for blacks (428) and Latinos (457) are significantly below those of whites (534) and Asians (598). The scores of black and Latino students are clustered towards the bottom of the distribution, while white scores are relatively normally distributed, and Asians are clustered at the top" [57]. This disparity is often attributed to racial and class inequalities encountered early in life, and presenting persistent obstacles to upward mobility and opportunity.

Technical bias. This type of bias arises from technical constraints or considerations, such as the screen size or a ranking's inherent position bias—the geometric drop in visibility for items at lower ranks compared to those at higher ranks. Position bias arises because in Western cultures we read from top to bottom, and from left to right, and so items appearing in the top-left corner of the screen attract more attention [7]. A practical implication of position bias in rankings that do not admit ties is that, even if two items are equally suitable for a searcher, only one of them can be placed above the other in a ranking, suggesting to the searcher that it is better and should be prioritized.

Note that, as all rankings carry an inherent position bias, any method that produces rankings with equalized candidate visibility implicitly addresses this technical bias. However, we will only assign a method to technical bias mitigation, if the article is explicitly concerned with it, such as [11].

Emergent bias. This type of bias arises in a context of use and may be present if a system was designed with different users in mind or when societal concepts shift over time. In the context of ranking and recommendation it arises most notably because searchers tend to trust the systems to indeed show them the most suitable items at the top positions [53], which in turn shapes a searcher's idea of a satisfactory answer for their search. These feedback loops can create a "the-winner-takes-it-all" situation in which consumers increasingly prefer one majority product over everything else.

3.3 Mitigation Objectives

3.3.1 Worldviews. Friedler et al. [34] reflect on the impossibility of a purely objective interpretation of algorithmic fairness (in the sense of a lack of bias): "In order to make fairness mathematically precise, we tease out the difference between beliefs and mechanisms to make clear what aspects of this debate are opinions and which choices and policies logically follow from those beliefs." They model the decision pipeline of a task as a sequence of mappings between three metric spaces: construct space (CS), observed space (OS), and decision space (DS), and define worldviews (belief systems) as assumptions about the properties of these mappings.

The spaces and the mappings between them are illustrated in Figure 5 for the college admissions example. Individuals are represented by points. CS represents the "true" unobservable properties of an individual (e.g., intelligence and grit), while OS represents the properties that we can measure

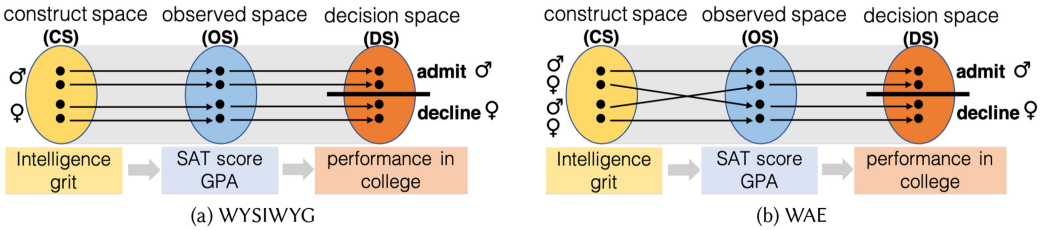

Fig. 5. An illustration of the worldviews from Frieder et al. [34]: "What you see is what you get" (WYSIWYG) vs. "We are all equal" (WAE). WYSIWYG assumes that the mapping from the construct space (*CS*) to the observed space (*OS*) shows very low distortion. In contrast, WAE assumes that the mapping from *CS* to *OS* distorts the structure of the groups in *CS*, leading to structural bias.

(e.g., SAT score as a proxy for intelligence, high school GPA as a proxy for grit) and serves as the feature space of an algorithmic ranker. An observation process $g(p) = \hat{p}$ maps from an individual $p \in CS$ to an entity $\hat{p} \in OS$. An example of such a process is an SAT test. The decision space *DS* maps from *OS* to a metric space of decisions, which for rankings represent the degree of relevance of an entity $\hat{p}$ by placing it at a particular position in the ranking.

Note that the mappings between the spaces are prone to distortions, of which those that map from *CS* to either *OS* or *DS* are by definition unobservable. Because the properties of these mapping cannot be independently verified, a belief system has to be postulated. Friedler et al. [34] describe two extreme cases: WYSIWYG ("what you see is what you get") and WAE ("we are all equal"). The former assumes that *CS* and *OS* are essentially the same and any distortion between the two is at most ϵ. The latter assumes that any differences between the utility distributions of different groups are due to an erroneous or biased observation process g. In our college admissions example this would mean that any differences in the GPA or IQ distributions across different groups are solely caused by biased school systems and IQ tests. It is also assumed that g shows different biases across groups, to which the authors refer as *group skew*.

The authors further define different terms from the Fairness, Accountability, Transparency, and Ethics (FATE) literature in terms of the underlying group skew. Their *fairness* definition is inspired by Dwork et al. [27] and says that items that are close in construct space shall also be close in decision space, which is widely known as individual fairness: similar individuals should receive similar outcomes. Group fairness, however, is defined indirectly through the terms *direct discrimination* and *non-discrimination*, requiring that an individual's treatment should not depend on their group membership. More formally, direct discrimination is absent if the group skew of a mapping between *OS* and *DS* is less than ϵ. Non-discrimination is present, if the group skew of a mapping between *CS* and *DS* is less than ϵ. Note that the last definition requires a choice of world view beforehand in order to be evaluated. If WYSIWYG is chosen, group fairness is given as soon as there is no direct discrimination, because $CS \approx OS$.

We will classify the investigated algorithms in terms of which worldview they choose and which of the three terms (fairness, direct discrimination, non-discrimination) they aim to optimize.

When categorizing surveyed methods with respect to worldview, we consider whether their fairness objective aims for equality of outcome or equality of treatment. If the goal of a method is to achieve equality of outcome, and if it is asserted that OS is not trustworthy because of biased or erroneous distortion g between CS and OS, then we consider this method to fall under the WAE worldview. If, on the other hand, the goal is to achieve equality of treatment and it is asserted that the mapping between CS and OS shows low distortion, then the method falls under the WYSIWYG worldview.

3.3.2 Equality of Opportunity. Equality of Opportunity (EO) is a philosophical doctrine that aims to remove morally irrelevant and arbitrary barriers to the attainment of desirable positions. Heidari et al. [39] show an application of equality of opportunity (EO) frameworks to algorithmic fairness: "At a high level, in these models an individual's outcome/position is assumed to be affected by two main factors: his/her circumstance c and effort e. Circumstance c is meant to capture all factors that are deemed irrelevant, or for which the individual should not be held morally accountable; for instance, c could specify the socio-economic status they were born into. Effort e captures all accountability factors—those that can morally justify inequality." Several conceptions of EO have been proposed, differing in what features they consider morally relevant, and in how the relationship between circumstance and effort is modeled.

Formal EO considers a competition to be fair when candidates are evaluated on the basis of their qualifications, and the most qualified candidate wins. This view rejects any qualifications that are irrelevant, such as hereditary privileges or social status, but it makes no attempt to correct for arbitrary privileges and disadvantages leading up to the competition that can lead to disparities in qualifications at the time of competition. Formal EO is typically understood in the algorithmic fairness literature as fairness-through-blindness—disallowing the direct impact from sensitive attributes (e.g., gender and race) on the outcome but allowing them to impact the outcome through proxies.

Limiting formal EO to fairness through blindness has been challenged in recent work by Khan et al. [40], who argue for a broader interpretation: "For example, think of the SAT as a predictor of college success: when students can afford to do a lot of test prep, scores are an inflated reflection of students' college potential. When students do not have access to test prep, the SAT underestimates students' college potential. The SAT systematically overestimates more privileged students, while systematically underestimating less privileged students. The test's validity as a predictor of college potential varies across groups. That is also a violation of formal EO. After all, in the college admissions contest, applicants should only be judged by "college-relevant" qualifications—but this test's accuracy as a yardstick for college potential varies with students' irrelevant privilege'." *Formal-plus EO*, due to Fishkin [32], addresses this important shortcoming of formal EO, capturing the desideratum that test performance should not skew along the lines of morally irrelevant factors. Tests that satisfy formal-plus EO include those that aim to balance error rates [41], as well as equalized odds [38].

Substantive EO doctrines take a broader view of Equal Opportunity—one that is not limited to fair competitions. Instead, they consider whether people have comparable opportunities over the course of a lifetime, including crucial developmental opportunities such as access to education. In order to make such a determination, substantive doctrines attempt to mitigate the effect of morally arbitrary factors such as gender, race, and socio-economic status, on people's relevant qualifications, which are the basis for attaining desirable positions. Importantly, in contrast to formal and formal plus EO that focus on the *current competition*, substantive EO aims to make people's *future prospects* comparable.

There are several prominent conceptions of substantive EO. *Luck-egalitarian EO* (see Dworkin [28] and Roemer [59]) would distribute outcomes after conditioning people's morally relevant qualification score on their morally irrelevant circumstances. Such an approach may, for example, rank individuals separately by group, and then take the specified number of top-ranked individuals from each list.

An alternative iterative approach to equalizing people's life chances could follow *Rawls' Fair EO*, and distribute outcomes in a way that improves the parity in people's future prospects of success, setting them up to be competitive in future competitions, even if it means "unfairness" in the outcomes of the current competition [56]. In this artcile, we will interpret fairness

interventions that attempt to model what an individuals' qualifications *would have looked like*, in a world where *equally talented people have equal prospects of success*, as Rawls's Fair EO. Once this has been satisfied, we can look more broadly at improving people's life prospects by applying the Difference Principle, which explicitly focuses on improving outcomes for the most disadvantaged (i.e., maximizing the minimum).

We will classify surveyed approaches with respect to the EO framework based on how their fairness definition compares individuals according to some qualifications (e.g., test scores, credit amount, and times of being arrested). However, such a mapping is elusive if a paper does not clearly state its assumptions about how morally arbitrary factors affect an individual's relevant qualifications. Note also that we explicitly map the *fairness definitions*, not the overall approaches. This is because many methods *combine* a fairness and a utility objective into a single optimization problem and, by doing so, lose a clear association with a particular framework. As a result, many of the methods we survey fall between the WAE and WYSIWYG worldviews, and do not cleanly map to a single EO category. Some are even designed to allow a continuous shift between the frameworks, by providing a tuning parameter [83].

Some authors [3, 39] categorize the *libertarian* view as an EO framework. According to this view, any information about an individual that was legally obtained can be used to make a decision. Because there is no attempt to equalize access to opportunity, this view corresponds to a narrow notion of procedural fairness, and we do not categorize it under EO [40]. If a fairness definition assumes that all individuals are comparable in all dimensions, as long as there is no gross violations of their privacy during the comparison, then we map this definition to the libertarian view.

Worldviews vs. Equality of Opportunity. The different worldviews [34] give us an intuitive way of thinking about the sufficiency of different EO doctrines. The WYSIWYG worldview assumes that there is no distortion between the construct space and the observed space, and, in such a setting, formal and formal-plus EO conceptions that focus on adjudicating outcomes fairly based on observable qualifications are sufficient. The WAE worldview, on the other hand, models structural bias that leads to the mis-measurement of qualifications of certain demographic groups. One way to correct for this is by adopting a formal-plus EO approach that attempts to eliminate skew in test performance between groups *at the point of competition*. Alternatively, the conditions modeled by WAE may be mitigated by interventions consistent with substantive EO, which seeks to equalize opportunities *over a lifetime* by modeling and controlling for *causes of the skew*.

3.3.3 Intersectional Discrimination. Intersectional Discrimination [23, 46] states that individuals who belong to several protected groups simultaneously (e.g., Black women) experience stronger discrimination compared to individuals who belong to a single protected group (e.g., White women or Black men), and that this disadvantage compounds more than additively. This effect has been demonstrated by numerous case studies, and by theoretical and empirical work [21, 25, 50, 63]. The most immediate interpretation for ranking is that, if fairness is taken to mean proportional representation among the top-k, then it is possible to achieve proportionality for each gender sub-group (e.g., men and women) and for each racial sub-group (e.g., Black and White), while still having inadequate representation for a sub-group defined by the intersection of both attributes (e.g., Black women).

Intersectional concerns also arise in more subtle ways. For example, Yang et al. [78] observed that when representation constraints are stated on individual attributes, like race and gender, and when the goal is to maximize score-based utility subject to these constraints, then a particular kind of unfairness can arise, namely, utility loss can be particularly severe in historically disadvantaged intersectional groups. When discussing specific technical methods, we will speak

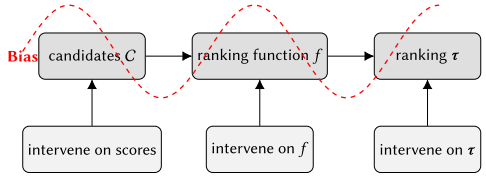

Fig. 6. Bias mitigation in score-based ranking: intervening on the score distribution of the candidates in C, on the ranking function f, or on the ranked outcome.

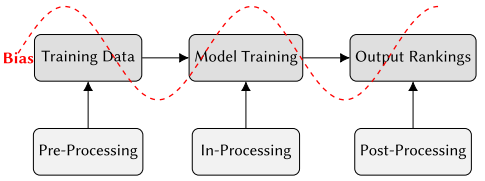

Fig. 7. Bias mitigation at different stages of supervised learning-to-rank: pre-processing, in-processing, and post-processing.

to whether they consider intersectional discrimination and, if so, which specific concerns they aim to address.

3.4 Mitigation Method

Score-based and supervised learning based rankers use different types of bias mitigation methods.

In score-based ranking, we categorize mitigation methods into those that intervene on the score distribution, or on the scoring function, or on the ranked outcome, as illustrated in Figure 6. Methods that *intervene on the score distribution* aim to mitigate disparities in candidate scores, either before these candidates are processed by an algorithmic ranker or during ranking. Methods that *intervene on the ranking function* identify a function that is similar to the input function but that produces a ranked outcome that meets the specified fairness criteria. Methods that *intervene on the ranked outcome* impose constraints to require a specific level of diversity or representation among the top-k as a set, or in every prefix of the top-k.

In supervised learning, we categorize mitigation methods into pre-processing, in-processing, and post-processing. This is illustrated in Figure 7, which is analogous to Figure 6. *Pre-processing* methods seek to mitigate discriminatory bias in training data, and have the advantage of early intervention on pre-existing bias. *In-processing* methods aim to learn a bias-free model. Finally, *post-processing* methods re-rank candidates in the output subject to given fairness constraints [37].

The advantage of post-processing methods in supervised learning is that they provide a guaranteed share of visibility for protected groups. In contrast, in-processing methods only consider fairness at training time and make no guarantees about fairness of the test set. However, post-processing methods may be subject to legal challenges because of due process concerns that may make it illegal to intervene at the decision stage (e.g., Ricci vs. DeStefano [70]). Thus, like all technical choices, the choice of whether to use a pre-, in-, or post-processing fairness-enhancing intervention is not purely technical, but must also consider the social and legal context of use of the algorithmic ranker.

4 DATASETS

Before diving into a description of the fair ranking methods, we present the experimental datasets used by them. We summarize the datasets in Table 2, where we highlight the following aspects: size (usually the number of candidates), sensitive attributes, scoring attributes, and the surveyed papers that use this dataset in their evaluation. We then briefly describe each dataset, and refer the reader to the description of each method for details about that dataset's use: score-based ranking in Section 5 and supervised learning and recommender systems in the second part of this survey. All datasets are publicly available under the referenced links unless otherwise indicated.

The papers surveyed here rarely substantiate their choice of an experimental dataset, other than that by the fact that the dataset was available, and that items in it have scores on which to rank.

Table 2. Experimental Datasets Used in the Surveyed Papers

Dataset	Size	Sensitive attrs	Score	Used in
AirBnB [1]	10,201 houses	gender of host	rating, price	[11, 43]
COMPAS [55]	7,214 people	gender, race	risk scores	[4, 80, 81]
CS departments [24]	51 departments	department size, geographic region	number of publications in different CS areas	[78]
DOT [52]	1.3 million flights	airline name	departure delay, arrival delay, taxi-in time	[4]
Engineering students [69]	5 queries, 650 students per query	gender, high school type	academic performance after first year	[82]
Forbes richest U.S. [33]	400 people	gender	net worth	[68]
German credit [58]	1,000 people	gender, age	credit amount, duration	[65, 75, 80, 81]
IIT-JEE [71]	384,977 students	birth category, gender, disability status	test scores	[15]
LSAC [62]	21,792 students	gender, race	LSAT scores	[83]
MEPS [51]	15,675 people	gender, race, age	number of trips requiring medical care	[78]
NASA astronauts [49]	357 astronauts	major in college	flight hours	[68]
Pantheon [19]	11,341 people	occupation	popularity of Wikipedia page	[68]
SAT [60]	1.6M students	gender	SAT score	[81]
StackExchange [67]	253,000 queries 6M documents	domains	document relevance	[11]
SSORC [61]	8,975,360 papers	gender of authors (inferred)	number of citations	[15]
W3C experts [73]	60 queries, 200 experts per query	gender	probability of being an expert	[82]
XING [76]	40 candidates	gender	years of experience, education	[43, 81]
Yahoo LTR [77]	26,927 queries 638,794 documents	N/A	relevance	[65]
Yow news [85]	unknown	source of news	relevance	[64]

Both of these reasons can be seen as purely technical (or even syntactic) rather than conceptual. Unfortunately little explicit attention has been paid to explaining whether a particular dataset was collected with a ranking task in mind, and *why* it is deemed appropriate for the specific fairness definition, that is, whether and to what extent the task for which the dataset was collected or can plausibly be used exhibits unfairness of the kind that the proposed fairness definition is designed to address. We see this as an important limitation of empirical studies in fairness in ranking and, more generally, in algorithmic fairness research, and posit that the use of a dataset must be explicitly justified.

AirBnB [1]. This dataset consists of 10,201 house listings from three major cities: Hong Kong (4,529 items), Boston (3,944 items), and Geneva (1,728 items). The gender of the hosts is used as the sensitive attribute, and the ranking score is computed as the ratio of the rating and the price.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [55]. This dataset is derived based on a recidivism risk assessment tool called COMPAS. The dataset contains the COMPAS scores from the Broward County Sheriff's Office in Florida in 2013 and 2014,

and the profile of each person's criminal history collected by ProPublica [2]. In total there are 7,214 data points, with sensitive attributes gender and race.

CS Department Rankings [24]. This dataset contains information about 51 computer science departments in the U.S. The methods in [78, 79] use the number of publications as the ranking score. Two categorical attributes are treated as sensitive: department size (with values "large" and "small") and geographic area (with values "North East", "West", "Middle West", "South Center", and "South Atlantic").

DOT (Department of Transportation) [52]. This dataset consists of about 1.3 million records of flights conducted by 14 U.S. airlines in the first three months of 2016. The dataset was collected by Asudeh et al. [4] from the flight on-time database that is published by the U.S. Department of Transportation. Three scoring attributes are used in [4]: departure delay, arrival delay, and taxi-in time. The name of the airline conducting the flight is treated as the sensitive attribute.

Engineering Students [69]. This dataset contains the results of a Chilean university admissions test from applicants to a large engineering school in five consecutive years. The task in [82] is to predict the students' academic performance after the first year based on their admissions test results and school grades. The sensitive attributes are gender and whether the applicants graduated from a private or public high school. This dataset is only accessible upon request, see referenced link for details.

Forbes Richest Americans [33]. This dataset consists of 400 individuals from the 2016 Forbes US Richest list,[1] ranked by their net worth. Gender is the sensitive attribute, with 27 female vs. 373 male individuals in the dataset.

German Credit [58]. This dataset, hosted by the UCI machine learning repository [44], contains financial information of 1,000 individuals, and is associated with a binary classification task that predicts whether an individual's credit is good or bad. The sensitive attributes are gender and age, where age is categorized into younger or older based on a threshold (25 or 35 years old is variably used as the threshold). Attributes credit amount and duration (how long an individual has had a line of credit) have been used as scoring attributes in fair ranking papers.

IIT-JEE (The Joint Entrance Exam of Indian Institutes of Technology) [71]. This dataset consists of scores of 384,977 students in the Mathematics, Physics, and Chemistry sections of IIT-JEE 2009, along with their gender, birth category (see [9]), disability status, and zip code. The students are scored on a scale from −35 to +160 points in all three sections, with an average total score of +28.36, a maximum score of +424, and a minimum score of −86.

LSAC [62]. This dataset consists of a U.S. national longitudinal bar exam passage data gathered from the class that started law school in Fall 1991. Data is provided by the students, their law schools, and state boards of bar examiners over a 5-year period [74]. The dataset consists of 21,791 students, with the sensitive attributes sex and race. Rankings are produced based on LSAT scores.

MEPS (Medical Expenditure Panel Survey) [51]. This dataset consists of 15,675 people and their information regarding the amount of health expenditures [18, 22]. The sensitive attributes are gender, race, and age of each individual, where age is categorized into younger or older based on a threshold (35 years old) in [78, 79]. The ranking score is based on utilization, defined by the IBM AI Fairness 360 toolkit [10] as the total number of trips requiring medical care. Utilization

[1] https://www.forbes.com/forbes-400/list/.

is computed as the sum of the number of office-based visits, the number of outpatient visits, the number of ER visits, the number of inpatient nights, and the number of home health visits.

NASA Astronauts [49]. This dataset consists of 357 astronauts with their demographic information. The method in [68] ranks this dataset by the number of space flight hours, and assigns individuals to categories based on their undergraduate major, treating is a the sensitive attribute. A total of 83 majors are represented in the dataset, the 9 most frequent are assigned to their individual categories—Physics (35), Aerospace Engineering (33), Mechanical Engineering (30), and so on, and the remaining 141 individuals are combined into the category "Other", resulting in 10 groups.

SAT [60]. This dataset contains about 1.6 million data points, in which the score column corresponds to an individual's results in the US SAT in 2014 [72]. The sensitive attribute is gender.

SSORC [61]. The Semantic Scholar Open Research Corpus contains the meta-data of 46,947,044 published research papers in computer science, neuroscience, and biomedicine from 1936 to 2019 on Semantic Scholar. The meta-data for each paper includes the list of authors of the paper, the year of publication, the list of papers citing it, and the journal of publication, along with other details. The sensitive attribute is the gender of the authors, collected by Celis et al. [15]. The ranking score is the number of citations of each paper.

StackExchange [67]. This dataset contains a query log and a document collection using the data from the Stack Exchange Q&A community (dump as of 13-06-2016) [12]. It consists of about 6 million posts inside the type "Question" or "Answer" in 142 diverse sub-forums (e.g., Astronomy, Security, Christianity, Politics, Parenting, and Travel). The questions are translated into about 253,000 queries, and the respective answers serve as the documents for the queries. The sensitive attribute is the query domain.

W3C experts [73]. The task behind this dataset corresponds to a search of experts for a given topic based on a corpus of e-mails written by possible candidates. The sensitive attribute is the gender of the expert. The experimental setup in [82] investigates situations in which bias is unrelated to relevance: expertise has been judged correctly, but ties have been broken in favor to the privileged group (i.e., all male experts are followed by all female experts, followed by all male non-experts, followed finally by all female non-experts).

Xing [76]. This dataset was collected by Zehlike et al. [81] from a German online job market website.[2] The authors collected the top-40 profiles returned for 54 queries, and computed an ad-hoc score based on educational features, job experience, and profile popularity. The sensitive attribute is gender, which was inferred based on the first name associated with the profile and the profile picture, when available. Items are ranked based on an ad-hoc score.

Yahoo! LTR [77]. This dataset consists of 19,944 training queries and 6,983 test set queries. Each query has a variable sized candidate set of documents that needs to be ranked. There are 473,134 training and 165,660 test documents. The query-document pairs are represented by a 700-dimensional feature vector. For supervision, each query-document pair is assigned an integer relevance judgments from 0 (bad) to 4 (perfect). The dataset is used to evaluate the effectiveness of Learning to Rank methods in [65], thus no sensitive attribute is specified.

[2] https://www.xing.com.

Table 3. Classification of Score-based Ranking Methods According to the Frameworks in Section 3

Method	Group structure	Bias	Worldview	EO	Intersectional
Rank-aware proportional representation [80]	one binary sensitive attr.	pre-existing	WAE	luck-egalitarian	no
Constrained ranking maximization [16]	multiple sensitive attrs.; multinary; handled independently	pre-existing	WAE	luck-egalitarian (1 sensitive attr. only)	no
Balanced diverse ranking [78]	multiple sensitive attrs.; multinary; handled independently	pre-existing; technical	WAE	luck-egalitarian	yes
Diverse k-choice secretary [68]	one multinary sensitive attr.	pre-existing	WAE	luck-egalitarian	no
Utility of selection with implicit bias [41]	one binary sensitive attr.	pre-existing; implicit	WAE	N/A	no
Utility of ranking with implicit bias [15]	multiple sensitive attrs.; multinary; handled independently	pre-existing; implicit	WAE	N/A	yes
Causal intersectionally fair ranking [79]	multiple sensitive attrs.; multinary; handled independently	pre-existing	WAE	Rawlsian	yes
Designing fair ranking functions [4]	any	pre-existing	any	any	yes

Yow News [85]. This dataset contains explicit and implicit feedback from a set of users for news articles in the "people" topic produced by Yow [86]. The ranking score is the explicitly given relevance field. The source of news is treated as the sensitive attribute.

5 SCORE-BASED RANKING

In this section, we present several methods for fairness in score-based ranking. Rather than giving a purely technical comparison, we re-iterate that the choice of a method should be based on assumptions about the nature of unfairness, and on the fundamental modeling choices. Table 3 summarizes the methods presented in this section according to the frameworks of Section 3. Additionally, every technical methods is placed on the mind map in Figure 4, to give a visual summary and as a means to compare the methods.

Recall that, in score-based ranking we categorize mitigation methods into those that intervene on the ranking process, on the score distribution, or on the scoring function. In Section 5.1, we describe methods that *intervene on the ranked outcome* by ensuring proportional representation across groups. Next, in Section 5.2, we discuss several methods that formulate *fairness* and *coverage-based diversity* constraints by specifying bounds on the number of candidates from groups of interest to be present in prefixes of a ranked list. These methods also intervene on the ranked outcome. Then, in Section 5.3, we describe methods that *intervene on the score distributions*. Finally, in Section 5.4, we present a method that treats the fairness objective as a black-box and proposes a geometric interpretation of score-based ranking to reach the objective by *intervening on the ranking function*.

5.1 Intervening on the Ranked Outcome: Rank-aware Proportional Representation

To the best of our knowledge, Yang and Stoyanovich [80] were the first to formalize rank-aware fairness, under the assumption that the scores based on which the ranking is produced encode pre-existing bias.

Consider a ranking in which candidates are assigned to one of two groups, $\mathcal{G}_1$ or $\mathcal{G}_2$, according to a single binary sensitive attribute (e.g., binary gender), and with one of these groups, $\mathcal{G}_1$, corresponding to the protected group (e.g., the female gender). The fairness measures proposed in this article are based on the following intuition: Because it is more beneficial for an item to be ranked higher, it is also more important to achieve proportional representation at higher ranks. The idea, then, is to take several well-known proportional representation measures and to make them *rank-aware*, by placing them within a framework that applies position-based discounts.

Fairness definition and problem formalization. Recall from Section 2 that position-based discounting is commonly used to quantify utility (Equation (3)) or prediction accuracy in a ranking that we will cover in the second part of this survey. In a similar vein, the use of position-based discounting in Yang and Stoyanovich [80] is a natural way to make set-wise proportional representation requirements rank-aware. Specifically, the idea is to consider a series of prefixes of a ranking $\boldsymbol{\tau}$, for $k = 10, 20, \ldots$, to treat each top-k prefix $\boldsymbol{\tau}_{1\ldots k}$ as a set, to compute *statistical parity* at top-k, and to compare that value to the proportion of the protected group in the entire ranking. (Naturally, perfect statistical parity is achieved when $k = n$.) The values computed at each cut-off point are summed up with a position-based discount. Based on this idea, the authors propose three fairness measures that differ in the specific interpretation of statistical parity: normalized discounted difference (rND), ratio (rRD), and KL-divergence (rKL).

Normalized discounted difference (rND) (Equation (7)) computes the difference between the proportions of the protected group $\mathcal{G}_1$ at the top-k and in the over-all population. Normalizer Z is computed as the highest possible value of rND for the given number of items n and protected group size $|\mathcal{G}_1|$.

$$\mathrm{rND}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{\log_2 k} \left(\frac{|\boldsymbol{\tau}_{1\ldots k} \cap \mathcal{G}_1|}{k} - \frac{|\mathcal{G}_1|}{n} \right). \tag{7}$$

Normalized discounted ratio (rRD) is defined analogously, as follows:

$$\mathrm{rRD}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{\log_2 k} \left(\frac{|\boldsymbol{\tau}_{1\ldots k} \cap \mathcal{G}_1|}{|\boldsymbol{\tau}_{1\ldots k} \cap \mathcal{G}_2|} - \frac{|\mathcal{G}_1|}{|\mathcal{G}_2|} \right). \tag{8}$$

When either the numerator or the denominator of a term in Equation (8) is 0, the value of the term is set to 0.

Finally, normalized discounted KL-divergence (rKL) uses Kullback–Leibler (KL) divergence to quantify the expectation of the logarithmic difference between two discrete probability distributions, P_k that quantifies the proportions in which groups are represented at the top-k:

$$P_k = \left(\frac{|\boldsymbol{\tau}_{1\ldots k} \cap \mathcal{G}_1|}{k}, \frac{|\boldsymbol{\tau}_{1\ldots k} \cap \mathcal{G}_2|}{k} \right), \tag{9}$$

and Q that quantifies the proportions in which groups are represented in the over-all ranking:

$$Q = \left(\frac{|\mathcal{G}_1|}{n}, \frac{|\mathcal{G}_2|}{n} \right). \tag{10}$$

KL-divergence between P_k and Q, denoted $D_{KL}(P_k||Q)$, is computed at every cut-off point k, and position-based discounting is applied as the values are compounded, with normalizer Z defined analogously as for rND:

$$\mathrm{rKL}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{\log_2 k} D_{KL}(P_k||Q). \tag{11}$$

candidate	A_1	A_2	Y
b	male	White	9
c	male	Black	8
d	female	White	7
e	male	White	6
f	female	White	5
k	female	White	4
l	male	White	3
o	female	Black	2

(a)

$\boldsymbol{\tau}$
b
c
d
e
f
k
l
o

(b)

$\tilde{\boldsymbol{\tau}}$
b
d
c
f
e
k
l
o

(c)

$\tilde{\boldsymbol{\tau}}$
b
c
e
l
d
f
k
o

(d)

$\tilde{\boldsymbol{\tau}}$
b
c
d
o
e
f
k
l

(e)

Fig. 8. **(a)** A set of applicants for college admissions C, with two binary sensitive attributes: A_1 (gender), with protected group $\mathcal{G}_\mathsf{F} = \{d, f, k, o\}$ and privileged group $\mathcal{G}_\mathsf{M} = \{b, c, e, l\}$; and A_2 (race), with protected group $\mathcal{G}_\mathsf{B} = \{c, o\}$ and privileged group $\mathcal{G}_\mathsf{W} = \{b, d, e, f, k, l\}$. Protected values of A_1 and A_2 are shown in orange, and privileged values in blue. **(b)** Ranking $\boldsymbol{\tau}$ sorts the applicants in descending order of their score Y, as shown in Figure 8(b), with male candidates appearing in higher proportion at the top ranks. **(c)** Ranking $\tilde{\boldsymbol{\tau}}$ for A_1 mixes candidates in approximately equal proportion by gender, with $p = 0.5$ in Algorithm 1, and is expected to achieve statistical parity for this attribute, since gender groups are represented in equal proportion in C. **(d)** Ranking $\tilde{\boldsymbol{\tau}}$ for A_1, with $p = 0$ in Algorithm 1, places all, or most, male applicants about the female applicants. **(e)** Ranking $\tilde{\boldsymbol{\tau}}$ for A_2 (race), with $p = 0.5$ in Algorithm 1, is expected to achieve equal representation by race at top ranks, but not statistical parity, since C is not balanced by race.

Note that, unlike rND and rRD, which are limited to a binary sensitive attribute, rKL can handle a multinary sensitive attribute and so is more flexible.

Experiments and observations. The authors evaluate the empirical behavior of the proposed fairness measures using real and synthetic datasets. Real datasets used are COMPAS [55] and German Credit [58], see Section 4 for details. Synthetic datasets are generated using an intuitive data generation procedure described below. This procedure was later used in the work of Zehlike et al. [81] and Wu et al. [75], and is of independent interest.

Recall that $\mathcal{G}_1$ represents the protected group and $\mathcal{G}_2$ represents the privileged group, and suppose for simplicity that each group constituted one half of the candidates C. An example is given in Figure 8(a), in which C contains eight candidates, four female ($\mathcal{G}_1$) and four male ($\mathcal{G}_2$). The data generation procedure, presented in Algorithm 1, takes two inputs: a ranking $\boldsymbol{\tau}$ of C and a "fairness probability" p, and it produces a ranking $\tilde{\boldsymbol{\tau}}$. The input ranking $\boldsymbol{\tau}$ is assumed to be generated by the vendor according to their usual process (e.g., based on candidate scores, as in Figure 8(b)). Algorithm 1 splits up $\boldsymbol{\tau}$ into two rankings: $\boldsymbol{\tau}_1$ of candidates in $\mathcal{G}_1$ and $\boldsymbol{\tau}_2$ of candidates in $\mathcal{G}_2$. It then repeatedly considers pairs of candidates at the top of the lists, $\boldsymbol{\tau}_1(1)$ and $\boldsymbol{\tau}_2(1)$, and decides which of these should be ranked above the other, selecting $\boldsymbol{\tau}_1(1)$ with probability p and $\boldsymbol{\tau}_2(1)$ with probability $1 - p$, and appending the selected candidate to $\tilde{(\boldsymbol{\tau})}$.

The parameter p specifies the relative preference between candidates in $\mathcal{G}_1$ and in $\mathcal{G}_2$. When $p = 0.5$, groups are mixed in approximately equal proportion for as long as there are items in both groups. This is illustrated in Figure 8(c) for the sensitive attribute A_1 (gender) and in Figure 8(e) for the sensitive attribute A_2 (race). When $p > 0.5$, members of the protected group $\mathcal{G}_1$ are preferred, and when $p < 0.5$ members of the privileged group $\mathcal{G}_2$ are preferred. In extreme cases, when $p = 0$, all (or most) members of $\mathcal{G}_2$ will be placed before any members of $\mathcal{G}_1$, as shown in Figure 8(d) for the sensitive attribute A_1 (gender). Note that candidates within a group always remain in the same relative order in $\tilde{\boldsymbol{\tau}}$ as in $\boldsymbol{\tau}$ (that is, there is *no reordering within a group*), but there may be *reordering between groups*.

ALGORITHM 1: FairGen

Require: Ranking τ, fairness probability p.
{Initialize the output ranking $\tilde{\tau}$.}
$\tilde{\tau} \leftarrow \emptyset$
$\tau_1 = \tau \cap \mathcal{G}_1$
$\tau_2 = \tau \cap \mathcal{G}_2$
while $(\tau_1 \neq \emptyset) \wedge (\tau_2 \neq \emptyset)$ **do**
  $r = random([0, 1])$
  {Append the next selected item to $\tilde{\tau}$}
  if $r < p$ **then**
    $\tilde{\tau} \leftarrow pop(\tau_1)$
  else
    $\tilde{\tau} \leftarrow pop(\tau_2)$
  end if
end while
{If any items remain in τ_1 or τ_2, append them to $\tilde{\tau}$}
$\tilde{\tau} \leftarrow \tau_1$
$\tilde{\tau} \leftarrow \tau_2$
return $\tilde{\tau}$

The proposed fairness measures—normalized discounted difference (rND), ratio (rRD), and KL-divergence (rKL)—are evaluated on rankings produced by Algorithm 1 with a range of values for p and with different relative proportions of $\mathcal{G}_1$ and $\mathcal{G}_2$ in C. The authors conclude that rKL is the most promising measure both because it is smooth and because it naturally generalized to multinary sensitive attributes.

This article also proposes a bias mitigation methodology, inspired by Zemel et al. [84], that integrates fairness objectives into an optimization framework that balance fairness against utility, with an experimental evaluation on the German credit dataset [58] (see details in Section 4).

Insights. The fairness definitions of this article aim to address pre-existing bias, per classification in Section 3.2.

Fairness is interpreted as equality of outcomes, suggesting an underlying assumption of WAE, per classification in Section 3.3. Assuming the existence of indirect discrimination in candidate scores (i.e., that the observation process between construct space *CS* and observable space *OS* is biased), the article aims to ensure a similar representation of groups in the ranked outcomes.

The approach is designed around conditioning qualification scores on morally-irrelevant circumstances: candidates are ranked according to score within a demographic group, and a ranked outcome is considered fair if the groups are mixed in equal proportion when the input is balanced, as in Figure 8(a) by A_1 (gender), or, more generally, when statistical parity is achieved at high ranks. Assuming that the goal of the competition is to make future prospects comparable, this is consistent with luck-egalitarian EO, per classification in Section 3.3. Figure 9 and Table 3 summarize our analysis.

5.2 Intervening on the Ranked Outcome: Diversity Constraints

In Section 5.1, we discussed how fairness measures that are based on (set-wise) proportional representation can be made rank-aware. The methods described in this section start with the observation that if the total number of candidates in C, and the number of candidates in each demographic group of interest, is available as input (i.e., that these quantities are known a priori or can be estimated), then any measure that aims to equalize or bound the difference in proportions can be equivalently re-formulated with the help of counts. Specifically, proportional representation constraints

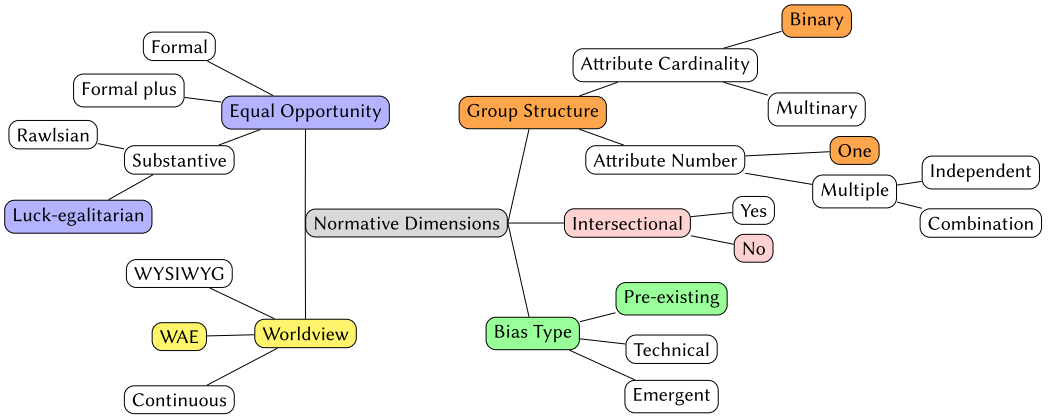

Fig. 9. Summary of the normative values encoded by Rank-aware proportional representation (Yang and Stoyanovich [80]).

and coverage-based diversity constraint [26] for *set selection tasks* can be expressed by specifying a lower-bound $L_k^{\mathcal{G}}$ and an upper-bound $U_k^{\mathcal{G}}$ on the representation of group $\mathcal{G} \subseteq C$ among the top-k set of a ranking. Such constraints can be formulated for one or several demographic groups of interest, and also for their intersections, and a score-based ranker can then optimize utility under such constraints. Generalizing beyond set selection, constraints $L_p^{\mathcal{G}}$ and $U_p^{\mathcal{G}}$ can be specified over every prefix of the top-k of a ranked list, with $p \in [k]$, or, more practically, at some specific cut-off points within the top-k.

Similarly to the methods of Section 5.1, the methods described in this section are designed to enforce fairness and diversity in the sense of representation. In contrast Section 5.1, these methods are designed to handle multiple sensitive attributes simultaneously—individually or in combination.

5.2.1 Celis et al. [16]. **Fairness definition and problem formalization.** The authors formulate the *constrained ranking maximization problem*: Consider a set of n candidates C, and the integer $k \ll n$, along with (1) the utility of placing a candidate in a particular position in the ranking, (2) the collection of sensitive attributes (e.g., gender, race, or disability status) that map candidates to groups $\boldsymbol{\mathcal{G}}$, and (3) a collection of lower-bound constraints $L_p^{\mathcal{G}}$ and upper-bound constraints $U_p^{\mathcal{G}}$ that, for each prefix $p \in [k]$ and for each group $\mathcal{G} \in \boldsymbol{\mathcal{G}}$, bound the number of candidates from that group that are allowed to appear in the top-p positions of the ranking. The goal is to output a ranking that maximizes overall utility with respect to the original utility metric, while respecting the constraints. Note that this problem formulation has the flexibility to explicitly associate a utility with an assignment of candidate $a \in C$ to rank position $j \in [k]$, and may already incorporate position-based discounting (per Equation (3)). However, for consistency and ease of exposition, we will assume that utility score Y is fixed per candidate.

An example of the constrained ranking maximization problem is given in Figure 10, where the goal is to select $k = 4$ candidates, with at least two of each gender ($L_4^{\mathtt{M}} = 2$, $L_4^{\mathtt{F}} = 2$) and at least one of each race ($L_4^{\mathtt{W}} = 1$, $L_4^{\mathtt{B}} = 1$, $L_4^{\mathtt{A}} = 1$) among the top-k, and with no further constraints on the prefixes of the top-k. (For convenience, we are referring to each groups by the first letter of the attribute value that defines it, such as `M` for `male` and `A` for `Asian`). Ranking $\boldsymbol{\tau}_1$ in Figure 10 is a ranked outcome of the top-4 candidates selected based on utility Y: two of them are male and two are female, and all are White. Applying diversity constraints on gender and race yields $\boldsymbol{\tau}_2$, a

candidate	A_1	A_2	Y
a	male	White	19
b	male	White	18
c	female	White	16
d	female	White	15
e	male	Black	11
f	male	Black	11
g	female	Black	10
h	female	Black	9
i	male	Asian	7
j	male	Asian	7
k	female	Asian	6
l	female	Asian	3

τ_1	τ_2	τ_3
a	a	a
b	b	c
c	g	e
d	k	k

Fig. 10. A set of applicants for college admissions C, with two sensitive attributes: A_1 (gender), with groups $\mathcal{G}_\mathsf{M} = \{a, b, e, f, i, j\}$ and $\mathcal{G}_\mathsf{F} = \{c, d, g, h, k, l\}$, and A_2 (race), with groups $\mathcal{G}_\mathsf{W} = \{a, b, c, d\}$, $\mathcal{G}_\mathsf{B} = \{e, f, g, h\}$, and $\mathcal{G}_\mathsf{A} = \{i, j, k, l\}$. Top-4 ranking $\boldsymbol{\tau}_1$ selects the highest-scoring candidates according to Y; all selected candidates are White, two of them are female and two are male. The utility of $\boldsymbol{\tau}_1$, computed as the sum of scores, is 68. Top-4 ranking $\boldsymbol{\tau}_2$ selects highest-scoring candidates subject to constraints to select at least two candidates of each gender, $L_4^\mathsf{M} = 2$, $L_4^\mathsf{F} = 2$, and at least one candidate of each race, $L_4^\mathsf{W} = 1$, $L_4^\mathsf{B} = 1$, $L_4^\mathsf{A} = 1$. The utility of $\boldsymbol{\tau}_1$, computed as the sum of scores, is 53. Top-4 ranking $\boldsymbol{\tau}_3$, subject to the same diversity constrains as $\boldsymbol{\tau}_2$, but additionally balancing utility loss within each group. This ranking has utility 52, and it returns the highest-scoring male, female, White, and Black candidates.

(a)

candidate	a	e	**b**	f	c	**d**
Y	7	4	8	5	9	3

(b)

candidate	a	e	**b**	**f**	c	d
Y	7	4	8	5	9	3

Fig. 11. An instance of the diverse k-choice secretary problem. A set of $n = 6$ college applicants C, are arriving for in-person interviews. The order of interviews is from left to right, with a arriving first, followed by e, and so on. A candidate's score Y is revealed when they are interviewed. C is partitioned into two groups based on (binary) gender, $\mathcal{G}_\mathsf{M} = \{a, b, c\}$ with $n_\mathsf{M} = 3$ candidates, and $\mathcal{G}_\mathsf{F} = \{d, e, f\}$ with $n_\mathsf{F} = 3$ candidates. The goal is to select $k = 2$ candidates, with one of every gender ($L_k^\mathsf{M} = 1$, $L_k^\mathsf{F} = 1$), and to maximize the expected sum of Y-scores subject to these diversity constraints. (a) using a common warm-up period yields candidates b and d, selecting the 2nd best male candidate but the lowest-scoring female candidate (b) separating warm-up per-group yields candidates b (as before) and f, the top-scoring female candidate.

ranking of the top-4 in Figure 10, selecting the top-scoring White male candidates a and b, and two lower-scoring female candidates, g and k. Computing total utility as the sum of scores of selected candidates (for simplicity), we observe that $U(\boldsymbol{\tau}_1) = 68$ and $U(\boldsymbol{\tau}_2) = 53$ in this example.

Note that the example in Figure 10 is deliberately constructed to highlight disparities in scores due to pre-existing bias on gender and race: all male candidates are ranked above all female candidates of a given race, and all Whites are ranked above all Black, who are in turn ranked above all Asians. For this reason, imposing diversity constraints leads to a substantial drop in score-utility of $\boldsymbol{\tau}_2$ in Figure 10.

In the example we constructed, diversity constraints are satisfiable. However, as was shown by Celis et al. [16], the constrained ranking maximization problem can be seen to generalize various NP-hard problems such as independent set, hypergraph matching and set packing, and so is hard in the general case. It turns out that even checking if there is a complete feasible ranking is NP-hard. The authors show that a special case of the problem, in which each candidate is assigned to (at most) one group, and so the assignment induces a partitioning on C, can be solved in polynomial time. In this case, diversity constraints can only be specified with respect to a single sensitive

attribute, which may be binary or multinary, and so can represent multiple sensitive attributes in combination (see discussion on group structure in Section 3.1).

Recall that the problem formulation allows to associate a utility with an assignment of candidate $a \in C$ to rank position $j \in [k]$. While the nature of these assignments can in principle be arbitrary, many reasonable utility metrics, including NDCG, Bradley-Terry [13], or Spearman's rho [66], are non-increasing with increasing rank position, and with decreasing utility score Y, which is intuitively interpreted to mean that, if $Y_a \geq Y_b$ then placing a above b cannot decrease the utility of the overall ranking. Such metrics are said to be monotone and to satisfy the Monge property. For this family of utility metrics, the authors propose an exact dynamic programming algorithm that solves the constrained ranking optimization problem in time polynomial in the number of candidates m and size of the selected set k, and exponential in the number of possible assignments of candidates to groups (typically the product of domain cardinalities of the sensitive attributes, $card(A_1) \times card(A_2) = 6$ in our example in Figure 10). The authors also propose approximation algorithms that allow violations of diversity constraints, and study the quality of these approximations.

Insights. The focus of this work is on the formal properties of the constrained ranking maximization problem, including its hardness and approximability under different assumptions about the sensitive attributes, the diversity constraints, and the properties of the utility metric. The paper does not include an experimental evaluation.

The paper states that "[...] left unchecked, the output of ranking algorithms can result in decreased diversity in the type of content presented, promote stereotypes, and polarize opinions." The goal of imposing diversity constrains is to counteract pre-existing bias. Furthermore, since these constraints enforce equality of outcomes, the method relates to the WAE worldview.

When the candidates are partitioned by a single sensitive attribute (either binary or multinary), the method will select the highest-scoring members of each group to satisfy diversity constraints. Under the assumption that the goal of the competition is to make future prospects comparable, this is consistent with substantive EO, and the mechanism is consistent with luck-egalitarian EO. However, when candidates are associated with two or more sensitive attributes, as is the case in the example in Figure 10, a single utility distribution is assumed, in the sense that a higher-scoring candidate will be preferred to a lower-scoring one irrespective of their group membership, whenever constraints permit. This was illustrated ranking $\boldsymbol{\tau}_2$ as shown in Figure 10, where the highest-scoring White and male candidates were selected among the top-k, but the highest-scoring female, Black, and Asian candidates were skipped. Based on this observation, the method falls short of satisfying the desiderata of EO for multiple sensitive attributes. Figure 12 and Table 3 summarize our analysis. We will elaborate on this specific concern in the next sub-section.

5.2.2 Yang et al. [78]. **Fairness definition and problem formalization.** The authors further investigate the constrained ranking maximization problem with two or more sensitive attributes, and observe that members of multiple historically disadvantaged groups may still be treated unfairly by this process. For an intuition, consider again Figure 10, and recall that the goal is to select the top-4 candidates, with at least two of each gender ($L_4^{\mathsf{M}} = 2, L_4^{\mathsf{F}} = 2$) and at least one of each race ($L_4^{\mathsf{W}} = 1, L_4^{\mathsf{B}} = 1, L_4^{\mathsf{A}} = 1$). Maximizing utility subject to these constraints being met yields, ranking $\boldsymbol{\tau}_2$ in Figure 10 that selects the best (according to score) White and male candidates a and b, but it does not select the best Black, Asian, or female candidates.

Our example was deliberately constructed to highlight the following: if some population groups have systematically lower scores, then it costs less to skip their best-scoring members in the name

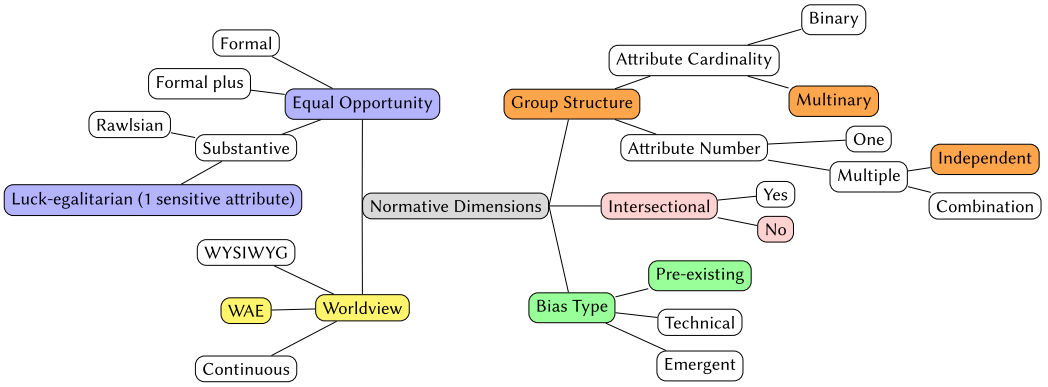

Fig. 12. Summary of the normative values encoded by Constrained ranking maximization (Celis et al. [16]).

of diversity. This runs contrary to the nature of the diversity objective, which is to equalize access to opportunity. This also represents unfairness, under the luck-egalitarian view. To see why, suppose that scores represent effort (e.g., how hard someone studied to do well on a test), and that we consider it important to reward effort. We may then take a relative view of effort, and assert that scores are more informative *within* a group than *across groups*. Taken together, this means that the best-scoring individuals from historically disadvantaged groups should have a chance to be selected among the top-k. Ranking $\boldsymbol{\tau}_3$ in Figure 10 represents a ranked outcome that gets closer to this objective; it presents $\boldsymbol{\tau}_3$, a top-4 ranking that contains the highest-scoring male, female, White and Black candidates.

Yang et al. [78] formalize this intuition by stating that, when multiple sensitive attributes (e.g., gender and race) are considered simultaneously, it is crucial to consider the utility loss that is incurred within each group, and to balance that loss across groups. The authors propose two measures to quantify in-group utility, IGF-Ratio and IGF-Aggregated, both taking on values from the range (0, 1), with 1 corresponding to perfect utility within a group (no loss), and with high loss corresponding to values close to 0. Both IGF-Ratio and IGF-Aggregated can be computed over the top-k as a set, or in rank-aware manner, by considering every prefix of length $p \in [k]$. In what follows, we will illustrate one of these measures, IGF-Ratio, taking the set interpretation for simplicity.

IGF-Ratio, quantifies the utility within a group (e.g., female or Black) by computing the ratio of the utility score of the highest-scoring skipped candidate from that group and the lowest-scoring selected candidate. Consider again ranking $\boldsymbol{\tau}_2$ in Figure 10. We compute $\text{IGF-Ratio}(\boldsymbol{\tau}_2, \mathcal{G}_M) = \text{IGF-Ratio}(\boldsymbol{\tau}_2, \mathcal{G}_W) = 1$, since the highest-scoring male and White candidates were selected. For the female, Black, and Asian groups, we compute $\text{IGF-Ratio}(\boldsymbol{\tau}_2, \mathcal{G}_F) = \frac{Y_g}{Y_c} = \frac{10}{16}$; $\text{IGF-Ratio}(\boldsymbol{\tau}_2, \mathcal{G}_B) = \frac{Y_g}{Y_e} = \frac{10}{11}$; and $\text{IGF-Ratio}(\boldsymbol{\tau}_2, \mathcal{G}_A) = \frac{Y_k}{Y_i} = \frac{6}{7}$.

IGF-Aggregated is based on similar intuition as IGF-Ratio, but rather than comparing the utility due to a pair of items for each group—one selected and one skipped—it compares the sum of scores of all items from a group up to a particular position with the sum of scores of all selected items from that group (up to the same position).

The authors go on to use IGF-Ratio and IGF-Aggregated to state that loss in these measures should be balanced across groups. They implement this requirement as an additional set of constraints, and formalize the induced optimization problem that (1) meets diversity constraints for each group, (2) balances utility loss across groups, and (3) maximizes over-all utility subject to (1) and (2), as integer linear programs.

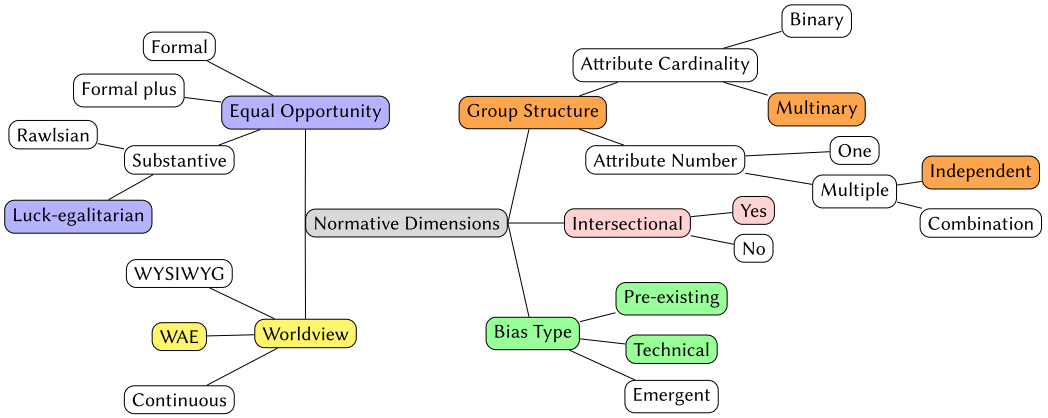

Fig. 13. Summary of the normative values encoded by Balanced diverse ranking (Yang et al. [78]).

Experiments and observations. The authors conduct experiments on two real datasets, CS departments [24] and MEPS [51] (see details in Section 4). They use these datasets to quantify the feasible trade-offs between, diversity, overall utility, and utility loss across groups. Furthermore, they show that utility loss can be balanced effectively, and that the over-all utility cost of such interventions is low.

Insights. Similarly to papers surveyed earlier in this section, the work of Yang et al. [78] aims to address pre-existing bias by equalizing outcomes, and so relates to the WAE worldview. Furthermore, because of an explicit focus on ensuring that the best-qualified candidates from each group have an opportunity to be selected, or to appear at higher ranks, this work conditions qualification score on morally irrelevant attributes (group membership), and so is firmly in the luck-egalitarian EO camp, under the assumption that the goal of the competition is to make individuals' future life prospects comparable. The main insight on which this work is based is that membership in multiple sensitive groups can lead to unfair treatment, and that the effects can be particularly pronounced for individuals who are multiply marginalized and who may, for example, be denied opportunity along the dimensions of both race and gender. This insight is surfacing an important dimension of intersectional discrimination in algorithmic rankers and is, to the best of our knowledge, the first approach in this area to have observed and proposed ways to counteract intersectional effects. Figure 13 and Table 3 summarize our analysis.

5.2.3 Stoyanovich et al. [68]. **Fairness definition and problem formalization.** The final method we discuss in this section aims to incorporate diversity constraints of the kind that were used by Celis et al. [16] and Yang et al. [78] into online set selection. This setting models a sequence of job or college admissions interviews: candidates arrive one-by-one and their qualification score Y is revealed at the time of the interview. Candidates are assumed to arrive in random order according to score, and their total number n is known or can be estimated. The decision maker must hire or to reject the candidate being considered as soon as their score Y is revealed, before advancing to the next candidate in the sequence.

The classic version of this problem, known as the Secretary problem [29, 45], aims to select a single candidate with the highest score Y. It was shown by Lindley [45] and by Dynkin [29] that the optimal hiring strategy is to interview $s = \lfloor \frac{n}{e} \rfloor$ candidates without making any offers (this is called the "warm-up period"), and make an offer to the first candidate whose score is better than the best score of the of the first s candidates (or accept the last candidate if no better candidate

is seen). This strategy yields the highest-scoring candidate with probability $\frac{1}{e}$, and is said to have "competitive ratio" e. Furthermore, this is the best such strategy for the secretary problem (i.e., with the highest competitive ratio) [31]. This problem has been extended by Babaioff et al. [6] to select k candidates, maximizing the expected sum of their scores. Stoyanovich et al. [68] postulate the diverse k-choice secretary problem that enriches the k-choice secretary problem of Babaioff et al. [6] with diversity constraints.

The diverse k-choice secretary problem is formalized as follows: In addition to a qualification score Y, each candidate is associated with one of $i \geq 2$ groups $\mathcal{G}$ based on the value of a single multinary sensitive attribute (e.g., gender, race, or disability status). Both the total number of candidates n, and the number of candidates in each group $n_1 \ldots n_i$, are known ahead of time or can be estimated. The goal of the decision maker is to select k candidates, maximizing the expected sum of their scores, subject to diversity constraints, stated in the form of per-group lower-bounds $L_k^{\mathcal{G}}$ and upper-bounds $U_k^{\mathcal{G}}$. Figure 11 gives an example: a set of $n = 6$ college applicants, of whom $n_{\mathsf{M}} = 3$ are male and $n_{\mathsf{F}} = 3$ are female, are being interviewed in the order shown in Figure 11, left-to-right. The admissions officer wishes to select $k = 2$ applicants, with one of each gender, specified by the lower-bound constraints $L_k^{\mathsf{M}} = 1$ and $L_k^{\mathsf{F}} = 1$.

The key idea in Stoyanovich et al. [68] is that, if score distributions are expected to differ between the groups, then separate warm-up periods should be conducted for each group to better estimate the scores of that group's desirable candidates. As illustrated in the outcome (a) in Figure 11, measuring the higher-scoring male candidates and the lower-scoring female candidates against the same (higher-scoring) standard will allow high-scoring male candidates to be selected. However, the female candidates selected in this way are those that happen to be at the end of the interview queue: they were chosen "at the last minute" to satisfy the diversity constraint. This is problematic for the reasons we outlined when discussing Yang et al. [78] earlier in this section—it withholds opportunity from the relatively better-qualified candidates of a historically disadvantaged group, and it can build bad precedent if the lesser-qualified candidates from that group are selected but do not perform well on the task. Outcome (b) in Figure 11 shows the result of a selection in which warm-up was conducted separately per group, yielding a higher-scoring female candidate.

The authors propose additional techniques to handle cases where the sum of the per-group lower bound is less than k, leaving the freedom to select high-scoring candidates from any group. Finally, they consider the case where a constant-size waiting list of candidates is allowed, showing that it can lead to higher-utility outcomes.

Experiments and observations. The experimental evaluation of the proposed algorithms for variants of the diverse k-choice secretary problems is conducted using three real datasets: Forbes richest Americans [33], NASA astronauts [49], and Pantheon [19] (see details in Section 4). Additional results on synthetic datasets are provided, to simulate differences in score distributions between groups. The evaluation on real datasets shows that the algorithms can select candidates that meet the desired diversity constraints while paying a very small cost in terms of the utility loss. The evaluation on synthetic datasets shows that if a difference in the observed scores is expected between groups, then these groups must be treated separately during processing. Otherwise, a solution may be derived that meets diversity constraints, but that results in lower utility for the disadvantaged groups.

Insights. This work focuses on pre-existing bias that exhibits itself through differences in expected scores between groups of candidates. Diversity constraints, and the mechanism used to enact them, aims to equalize outcomes across groups, and so this method clearly links to the WAE worldview. The core idea in this work is that effort, as represented by scores, should be seen as relative: scores are estimated per group, and individuals from a particular group are evaluated against that group's

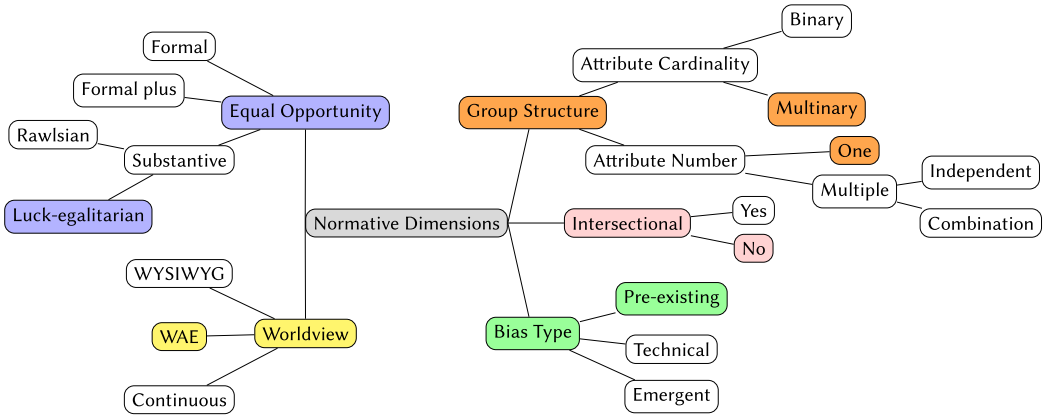

Fig. 14. Summary of the normative values encoded by Diverse k-choice secretary problem (Stoyanovich et al. [68]).

score threshold. Thus, under the assumption that the goal of the fainess intervention is to equalize opportunities over a lifetime, this method is consistent with luck-egalitarian EO. Figure 14 and Table 3 summarize our analysis.

5.3 Intervening on the Score Distribution

The methods in this sub-section work under the assumption that the scores on which candidates are ranked are subject to pre-existing bias, such that members of minority or historically disadvantaged groups have lower scores, and thus are ranked less favorably. The approach these methods take is based on correcting for the bias by adjusting the score distribution before it is given as input to a ranker.

5.3.1 Kleinberg and Raghavan [41] and Celis et al. [15]. **Problem formalization.** The papers discussed in this section study set selection and ranking in presence of *implicit bias*; they investigate under what conditions the utility of the selected set or the top-k would be improved by imposing representation constraints. Kleinberg and Raghavan [41] consider a score-based set selection task motivated by hiring, in which a set of n candidates C applies for an open job position, and some $k \ll n$ of them are selected as finalists to interview. The size of the selected set k is assumed to be a small constant, with the case $k = 2$ studied closely in the paper. Candidates in C belong to one of two groups, $\mathcal{G}_1$ or $\mathcal{G}_2$, according to a single binary sensitive attribute (e.g., binary gender), and with one of these groups, $\mathcal{G}_1$, corresponding to the protected group (e.g., the female gender). It is assumed that $\mathcal{G}_1$ constitutes a minority of the applicant pool, as quantified by the parameter $\alpha \in (0, 1]$, with $|\mathcal{G}_1| = \alpha \cdot |\mathcal{G}_2|$. Furthermore, it is assumed that the true qualification scores (called "potentials" in Kleinberg and Raghavan [41]) are drawn from the same score distribution for the candidates in both groups, and that this distribution follows the power law, parameterized by $\delta > 0$, such that $Pr[Y \geq t] = t^{-(1+\delta)}$.

Candidates are not hired according to their true qualification scores Y, but rather according to their perceived scores $\tilde{Y}$, which are, in turn, subject to *implicit bias*: hiring committee members "downgrade" the true scores of the candidates from $\mathcal{G}_1$ by dividing them by a factor $\beta > 1$.

The question being asked in this papers is: Under what conditions does including *a single candidate* from the protected group $\mathcal{G}_1$ among the k finalists improve the utility of the selected set according to the true score Y? (Utility is quantified as the sum of true scores of the selected candidates.) This intervention is known as the Rooney Rule [20], and while its goal is to improve

candidate	A (sex)	Y	$\tilde{Y}$
b	male	9	9
c	male	9	6
d	female	8	4
e	male	7	5
f	female	6	3
g	male	5	5

$\boldsymbol{\tau}$	$\tilde{\boldsymbol{\tau}}$
b	b
c	c
d	e
e	d
f	g
g	f

$\boldsymbol{\tau}^R$
b
d

Fig. 15. Consider a set of applicants for college admissions. Observed scores $\tilde{Y}$ of the applicants are affected by implicit bias: for the male candidates, $\tilde{Y} = Y$, while for the female candidates, $\tilde{Y} = Y/\beta$, with $\beta = 2$. Female candidates constitute a minority, with $\alpha = 1/2$ — there are 2 male candidates for each female one. Ranking $\boldsymbol{\tau}$ sorts candidates on their true (unobserved) scores Y; ranking $\tilde{\boldsymbol{\tau}}$ sorts them on scores $\tilde{Y}$ that are subject to implicit bias; ranking $\boldsymbol{\tau}^R$ applies to Rooney Rule to include the top-scoring female candidate among the top-2.

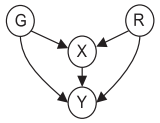

Fig. 16. A causal model that include sensitive attributes G (gender), R (race), utility score Y, and other covariates $\mathbf{X}$.

diversity in hiring, Kleinberg and Raghavan [41] study it explicitly from the point of view of utility rather than diversity or fairness. The requirement of including a single protected group candidate among the finalists is a basic coverage-based diversity requirement [26].

The authors study the problem under different settings of α (relative proportion of the minority group), β (bias factor), and δ (the parameter of the power law distribution of true scores). They find that, for every α, there exists a sufficiently small $\delta > 0$ for which the Rooney Rule will produce a set of k finalists with higher expected utility, compared to when candidates are selected according to their perceived—and biased—scores $\tilde{Y}$. Put another way, with a power law exponent $1 + \delta$ that is sufficient close to 1, it is a better strategy, *in terms of utility*, to commit one of the k offers to the candidates from group $\mathcal{G}_1$, even when k is as low as 2 and $\mathcal{G}_1$ forms an extremely small fraction of the population.

Figure 15 shows an example of the selection process, where the goal is to select $k = 2$ finalists from a pool of six, with two male candidates for each female candidate ($\alpha = 1/2$), and with females candidates being perceived as half as qualified as what their true score would suggest ($\beta = 2$). The Rooney Rule would select a top-scoring candidate from each gender group, leading to higher expected utility than if the top-two candidates were selected, both of them male.

The results of are extended by Celis et al. [15], who consider arbitrary utility distributions (beyond the power law) and support a richer group structure, including multiple sensitive attributes handled independently, with multinary domains. They show that, for any (assumed) distribution of utilities and any level of implicit bias, representation constraints can lead to optimal latent utility.

Experiments and insights. Kleinberg and Raghavan [41] give a tight characterization of the conditions on α, β, and δ, under which applying the Rooney Rule, with its most basic representation constraint, produces a positive change in expected utility. Proposed techniques can be used to estimate parameters of a biased decision-making process. The paper focuses on theoretical analysis and does not provide any experimental results.

Celis et al. [15] extend these results, and also include an experimental evaluation on the *IIT-JEE* dataset [71] (see Section 4 for details). These results give the flavor of the utility of the proposed

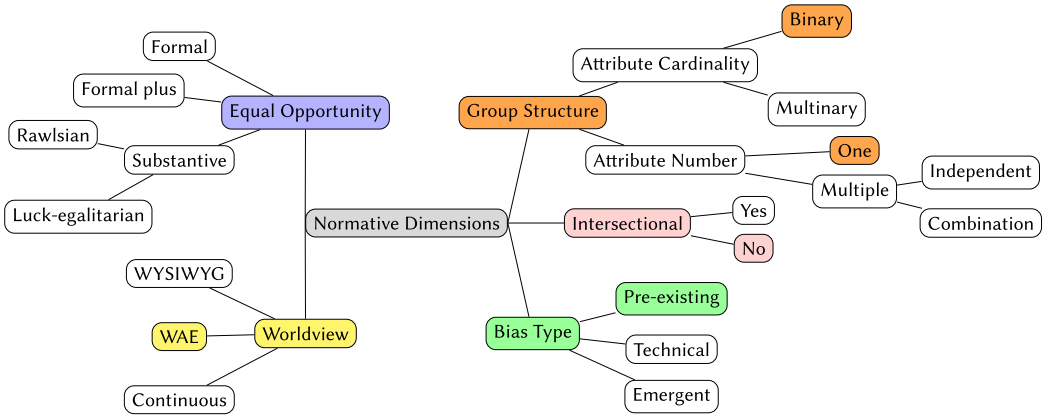

Fig. 17. Summary of the normative values encoded by Selection with implicit bias (Kleinberg and Raghavan [41]).

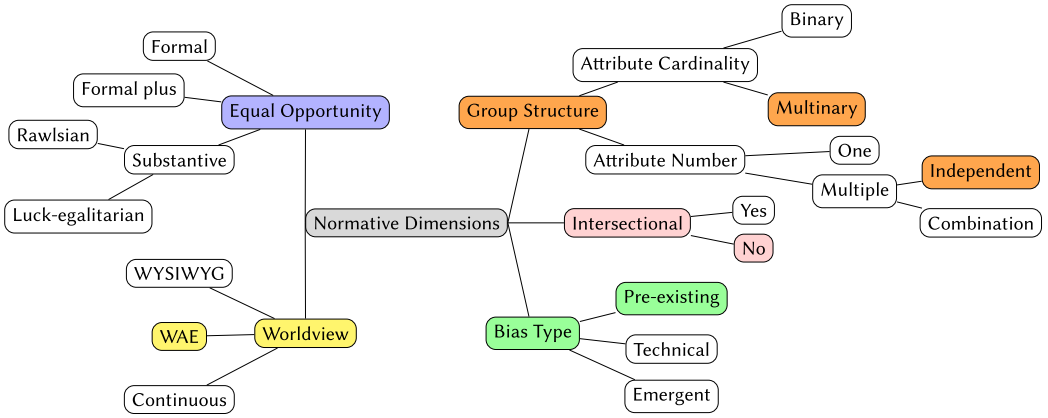

Fig. 18. Summary of the normative values encoded by Ranking with implicit bias (Celis et al. [15]).

intervention, although experimental evaluation is substantially more limited than the problem set-up warrants, focusing on a single binary protected attribute and leaving empirically unsubstantiated the claim that proposed approach generalizes to multiple sensitive attributes and handles intersectional discrimination.

Insights. Both papers consider utility rather than diversity or fairness, and so cannot be classified according to one of our EO frameworks. That said, the assumption made in the papers—that candidates' true (unobserved) qualifications are drawn from the same score distribution—is consistent with the WAE worldview. Figures 17 and 18 summarize our analysis for the above two methods [41] and [15], respectively. The summary can be also found in Table 3.

5.3.2 Yang et al. [79]. **Fairness definition and problem formalization.** The authors define *intersectional fairness for ranking* by modeling the causal effects of sensitive attributes on other variables, and then making rankers fairer by removing these effects. Their method, CIF-Rank, computes model-based counterfactuals to answer the question: "What would this person's data look like if they had (or had not) been a Black woman (for example)?" Counterfactual scores are

computed by treating every candidate as though they had belonged to one specific intersectional sub-group. Candidates are then ranked on counterfactual scores (for score-based rankers), or these scores are used to train a fair model (for rankers based on supervised learning).

Consider the hiring process of a moving company that has a dataset of applicants including their gender G, race R, weight-lifting test score X, and an overall qualification score Y by which job candidates are ranked. Figure 16 presents the **structural causal model** (**SCM**) that describes the data generation process. An SCM is a directed acyclic graph, where vertices represent (observed or latent) variables and edges indicate causal relationships from source to target vertices. The arrows pointing from G (gender) and R (race) directly to Y encode the effect of "direct" discrimination. Additionally, the SCM can encode indirect discrimination: note that G and R both impact Y through weight-lifting ability X, called a "mediator variable." A mediator may be designates as *resolving* with respect to a sensitive variable, which means that we allow that mediator to carry the effect from the sensitive variable to the outcome. For example, we may consider X as a resolving on the path from gender G to score Y. Alternatively, a mediator may be designates as *non-resolving*, which means that we consider the influence to be due to discrimination. For example, we may consider X as non-resolving on the path from race R to score Y.

The SCM, together with the information about which mediators are considered resolving, is given as input; it encodes the fairness objectives of the ranker. CIF-Rank will use the SCM to produce a ranking that is fair with respect to race, gender, and the intersectional sub-groups of these categories.

Let $\mathbf{A}$ denote the vector of sensitive attributes and let $\mathbf{a}$ denote a possible value. The counterfactual $Y_{\mathbf{A}\leftarrow\mathbf{a}'}$ is computed by replacing the observed value of $\mathbf{A}$ with $\mathbf{a}'$ and then propagating this change through the DAG: any directed descendant of $\mathbf{A}$ has its value changed by computing the expectation for the new value of $\mathbf{a}'$, and this operation is iterated until it reaches all the terminal nodes that are descendants of any of the sensitive attributes $\mathbf{A}$. If a mediator variable is non-resolving, then its value will be set to its counterfactual value in the process. If, however, it is designated as resolving, then we keep its observed value.

CIF-Rank considers a ranking $\hat{\boldsymbol{\tau}}$ is counterfactually fair if, for all possible x and pairs of vectors of actual and counterfactual sensitive attributes $a \neq a'$, respectively,

$$\begin{aligned}&\mathbb{P}(\hat{\boldsymbol{\tau}}(Y_{\mathbf{A}\leftarrow\mathbf{a}}(U)) = k \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a})\\ =\ &\mathbb{P}(\hat{\boldsymbol{\tau}}(Y_{\mathbf{A}\leftarrow\mathbf{a}'}(U)) = k \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a})\end{aligned}$$

for any rank k, and with suitably randomized tie-breaking.

The causal model can be used to compute counterfactual scores Y —the scores that would have been assigned to the individuals if they belonged to one particular sub-group defined by fixed values of R and G, while holding the weight lifting score X fixed in the resolving case—and then rank the candidates based on these scores. The moving company can then interview or hire the highly ranked candidates, and this process would satisfy a causal and intersectional definition of fairness that corresponds to the hiring manager's explicitly stated goals.

Experiments and observations. The authors evaluated the performance of CIF-Rank on several real and synthetic datasets, including *CSRankings*, *COMPAS*, and *MEPS* (see details in Section 4). Results on synthetic datasets are provided to simulate different structural assumptions of the underlying causal model. The evaluation is done on two types of ranking tasks: score-based and supervised learning. The evaluation of score-based ranking tasks on real and synthetic datasets shows that CIF-Rank can be flexibly applied to different scenarios, including ones with mediating variables and numerical sensitive attributes. Counterfactually fair rankings that are produced by CIF-Rank compare reasonably to intuitive expectations we may have about intersectional

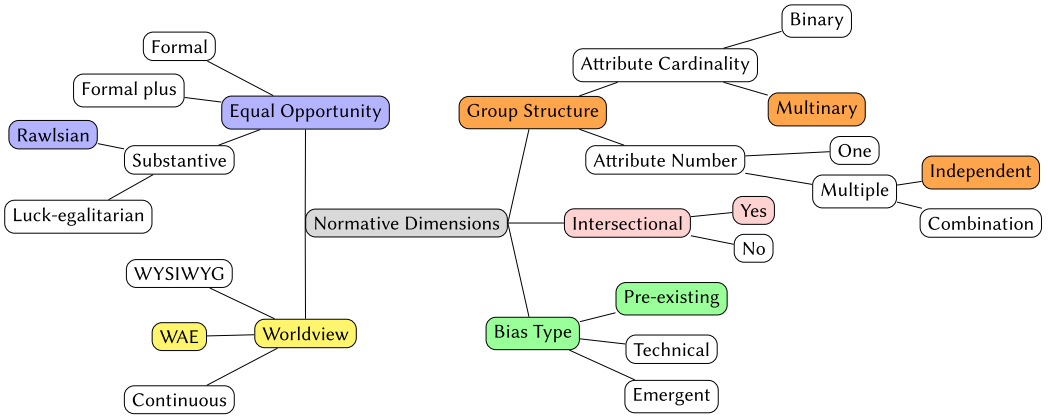

Fig. 19. Summary of the normative values encoded by Causal intersectionally fair ranking (Yang et al. [79]).

fairness for those examples, while paying a small cost in terms of the utility loss. The evaluation of rankers based on supervised learning on synthetic datasets shows that CIF-Rank can be used as a pre-processing fairness intervention to produce counterfactually fair training and test data.

Insights. CIF-Rank admits *multiple sensitive attributes*, and is specifically designed for intersectional concerns, and so is appropriate when it is important to account for potential discrimination along two or more features. The method supports *multinary sensitive attributes*, such as non-binary gender and ethnic group membership. The method is concerned with *pre-existing bias* that in turn leads to disparities in outcomes. The method focuses on *equality of outcome* and takes the *WAE* worldview. Under the assumption that the goal is to equalize opportunity over a lifetime, the method is consistent with *substantive EO*. It gives the decision maker the flexibility to specify which impacts of which sensitive attribute to mitigate, and which to allow to persist. This is done through the mediator mechanism. A mediator X may be considered resolving or not; this decision can be made separately for different sensitive attributes, and the relative strengths of causal influences of sensitive attributes on both X and Y can vary, creating potential for explanatory nuance. This method supports fairness interventions that attempt to model what an individuals' qualifications *would have looked like*, in a world where equally talented people have equal prospects of success. For this reason, we classify it as Rawls's Fair EO. Figure 19 and Table 3 summarize our analysis.

5.4 Intervening on the Ranking Function

To motivate the methods discussed in this section, let us return to our running example described in Section 1.1, and shown in Figure 1, and consider a college admissions officer who is designing a ranking scheme to evaluate a pool of applicants, each with several potentially relevant attributes. For simplicity, let us focus on two of these attributes, high school GPA X_1, and verbal SAT X_2, and assume that they are appropriately normalized and standardized. Suppose that our fairness criterion is that the admitted class comprise at least 40% women. The admissions officer may believe *a priori* that X_1 and X_2 should carry an approximately equal weight, computing the score of an applicant $a \in C$ as $f(a) = 0.5X_1 + 0.5X_2$, ranking the applicants, and returning the top 500 individuals. Upon inspection, it may be determined that an insufficient number of women is returned among the top-k: at least 200 were expected and only 150 were returned, violating the fairness constraint.

A possible mitigation is to identify an alternative scoring function $\tilde{f}$ that, when applied to C, meets the fairness constraint and is close to the original function f in terms of attribute weights,

thereby reflecting the admission officer's notion of quality. To arrive at such a function, the admissions officer would try a new scoring function, check whether the result meets the fairness criterion, and, if necessary, repeat. After a few cycles of such interaction, the admissions officer may choose $f(a) = 0.45X_1 + 0.55X_2$ as the final scoring function. The work of Asudeh et al. [4] automates this process; the authors use a combinatorial geometry approach to efficiently explore the search space and identify a fair scoring function $\tilde{f}$ in the neighborhood of f, if one exists.

Fairness definition and problem formalization. Let us assume that a dataset of candidates C is given, along with a linear ranking function f, specified by a weight vector $\vec{w}$. The goal is to find a ranking function $\tilde{f}$ that is both close to f in terms of the angular distance between the weight vectors of f and $\tilde{f}$, and fair according to a fairness oracle O.

The main technical contribution of the work is in establishing a correspondence between the space of linear ranking functions and the rankings of items from a given dataset C induced by these functions. This characterization is based on the notion of an *ordering exchange* that partitions the space of linear functions into disjoint regions. Intuitively, while there is an infinite number of linear ranking functions to explore, only those of them that change the relative order among some pair of items $a, b \in C$ need to be considered, because if a ranking is unchanged, then the fairness oracle O will not change its answer from *false* to *true*. Based on this observation, the authors develop exact algorithms to determine boundaries that partition the space into regions where the desired fairness constraint is satisfied, called satisfactory regions, and regions where the constraint is not satisfied. They also develop approximation algorithms to efficiently identify and index satisfactory regions, and introduce sampling heuristics for on-the-fly processing in cases where the size of C or the number of scoring attributes are large.

Experiments and observations. While the fairness model is general, the authors focus their experimental evaluation on proportional representation constraints that bound the number of items belonging to a particular group at the top-k, for some given value of k. Proposed methods are evaluated on the COMPAS [55] and DOT [52] datasets (see Section 4 for details), and with two sets of fairness measures: (1) proportional representation on a single multinary protected attribute and (2) proportional representation on multiple, possibly overlapping, protected attributes. They study both how intuitive the results are—how close a fair ranking function is to the original—and how efficiently results can be computed in this computationally challenging setting.

Insights. The fairness oracle O is treated as a black box: given a dataset C and a ranking function f, it returns *true* if the ranking of C by f meets fairness criteria and so is satisfactory, and returns *false* otherwise. The oracle is deterministic, and no further assumptions are made about the type of fairness criteria it encodes. Because of this black-box treatment of the fairness objective, the method makes no commitment to worldview (WYSWYG or WAE) or EO framework, and it is not restricted in terms of group structure: the number of sensitive attributes, their cardinality, and the method by which multiple sensitive attributes are handled. Despite this flexibility, the authors target their approach specifically at pre-existing bias. Figure 20 and Table 3 summarize our analysis.

6 SUMMARY OF PART I

This concludes Part I of the survey on fairness in ranking. In this part of the survey, we motivated fairness in ranking, presented notation, discusses several frameworks with respect to which we classify fair rankers, presented evaluation datasets, and, finally, dove deeply into fairness in score-based ranking.

In Part II of the survey we will build on Part I and present technical work on fairness in supervised learning to rank. We will also highlight some recent work on fairness in recommender

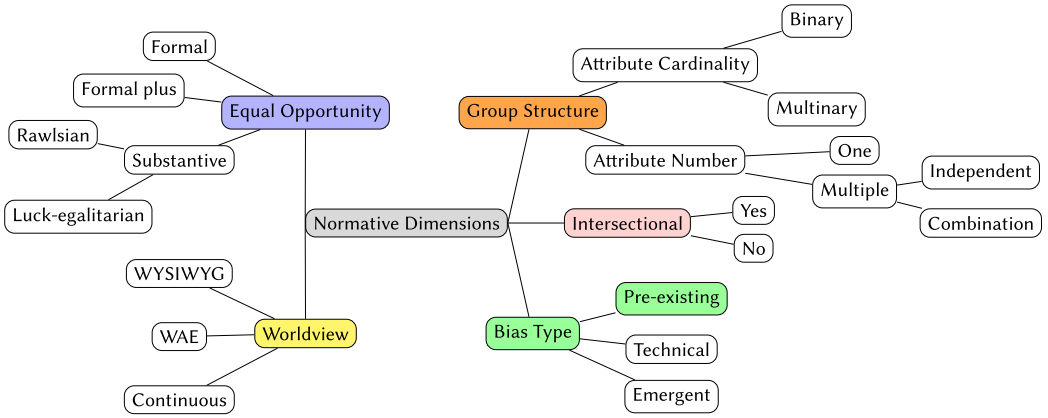

Fig. 20. Summary of the normative values encoded by Designing fair ranking functions (Asudeh et al. [4]).

systems and matching. Furthermore, we will discuss evaluation frameworks, present important directions of future work, and draw a set of recommendations on the evaluation of fair ranking methods.

ACKNOWLEDGMENTS

We are grateful to Falaah Arif Khan for her input on equality of opportunity (EO) frameworks, and on the mapping of specific methods to EO doctrines.

REFERENCES

[1] AirBnB. AirBnB. (????). https://insideairbnb.com.
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[3] Richard J. Arneson. 2018. Four conceptions of equal opportunity. *Economic Journal* 128, 612 (2018), 152–173.
[4] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*. ACM, 1259–1276.
[5] Abolfazl Asudeh and H. V. Jagadish. 2020. Fairly evaluating and scoring items in a data set. *Proceedings of the VLDB Endowment* 13, 12 (2020), 3445–3448. DOI: https://doi.org/10.14778/3415478.3415566
[6] Moshe Babaioff, Nicole Immorlica, David Kempe, and Robert Kleinberg. 2008. Online auctions and generalized secretary problems. *SIGecom Exchanges* 7, 2 (2008). DOI: https://doi.org/10.1145/1399589.1399596
[7] Ricardo Baeza-Yates. 2018. Bias on the web. *Communication of the ACM* 61, 6 (2018), 54–61. DOI: https://doi.org/10.1145/3209581
[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Nips Tutorial* 1 (2017), 2.
[9] Surender Baswana, P. P. Chakrabarti, Yashodhan Kanoria, Utkarsh Patange, and Sharat Chandran. 2019. Joint seat allocation 2018: An algorithmic perspective. arXiv:1904.06698. Retrieved from http://arxiv.org/abs/1904.06698.
[10] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943. Retrieved from http://arxiv.org/abs/1810.01943.
[11] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 405–414.
[12] Asia J. Biega, Rishiraj Saha Roy, and Gerhard Weikum. 2017. Privacy through solidarity: A user-utility-preserving framework to counter profiling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 675–684.
[13] Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[14] Carlos Castillo. 2019. Fairness and transparency in ranking. In *Proceedings of the ACM SIGIR Forum*, Vol. 52. ACM New York, NY, 64–71.
[15] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2020. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 369–380.
[16] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with fairness constraints. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
[17] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Communication of the ACM* 63, 5 (2020), 82–89. DOI: https://doi.org/10.1145/3376898
[18] Joel W. Cohen, Steven B. Cohen, and Jessica S. Banthin. 2009. The medical expenditure panel survey: A national information resource to support healthcare cost research and inform policy and practice. *Medical Care* (2009), S44–S50.
[19] MIT Collective Learning Group. Pantheon. (????). Retrieved from https://github.com/DataResponsibly/Datasets.
[20] Brian Collins. 2007. *NYU Law Review* 82, 3 (2007), 870.
[21] Patricia Hill Collins. 2002. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. routledge.
[22] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*. ACM.
[23] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43 (1990), 1241.
[24] CS Rankings. CSRankings: Computer Science Rankings. (????). Retrieved from https://csrankings.org.
[25] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. MIT Press.
[26] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in big data: A review. *Big Data* 5, 2 (2017), 73–84.
[27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.
[28] Ronald Dworkin. 1981. What is equality? part 1: Equality of welfare. *Philosophy and Public Affairs* 10, 3 (1981), 185–246. Retrieved from http://www.jstor.org/stable/2264894.
[29] E. B. Dynkin. 1963. The optimum choice of the instant for stopping a Markov process. *Sov. Math. Dokl.* 4 (1963), 627–629.
[30] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and discrimination in recommendation and retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 576–577. DOI: https://doi.org/10.1145/3298689.3346964
[31] Thomas S. Ferguson. 1989. Who solved the secretary problem? *Statistical Science* 4, 3 (Aug. 1989), 282–289. DOI: https://doi.org/10.1214/ss/1177012493
[32] Joseph Fishkin. 2014. *Bottlenecks: A New Theory of Equal Opportunity*. Oup Usa.
[33] Forbes. Forbes Richest Americans. (????). Retrieved from https://github.com/DataResponsibly/Datasets.
[34] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. arXiv:1609.07236. Retrieved from https://arxiv.org/abs/1609.07236.
[35] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information System* 14, 3 (1996), 330–347. DOI: https://doi.org/10.1145/230538.230561
[36] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. 2019. Fairness-Aware Neural Réyni Minimization for Continuous Features. arXiv:1911.04929. Retrieved from https://arxiv.org/abs/1911.04929.
[37] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2125–2126.
[38] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems*. 3315–3323.
[39] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 181–190.
[40] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2021. Translation tutorial: Fairness and friends. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.
[41] Jon Kleinberg and Manish Raghavan. 2018. Selection problems in the presence of implicit bias. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (Leibniz International Proceedings in Informatics)*, Anna R. Karlin (Ed.), Vol. 94. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 33:1–33:17. https://doi.org/10.4230/LIPIcs.ITCS.2018.33

[42] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, April 26–30, 2010*, Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti (Eds.). ACM, 571–580. DOI: https://doi.org/10.1145/1772690.1772749
[43] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *Proceedings of the 2019 IEEE 35th International Conference on Data Engineering*. IEEE, 1334–1345.
[44] M. Lichman. 2013. UCI Machine Learning Repository. (2013).
[45] D. V. Lindley. 1961. Dynamic programming and decision theory. *Journal of the Royal Statistical Society* 10, 1 (March 1961), 39–51.
[46] Timo Makkonen. 2002. Multiple, compound and intersectional discrimination: Bringing the experiences of the most marginalized to the fore. *Institute for Human Rights, Åbo Akademi University* (2002).
[47] Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. 2019. Fairness-aware learning for continuous attributes and treatments. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 4382–4391. Retrieved from https://proceedings.mlr.press/v97/mary19a.html.
[48] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), 115:1–115:35. DOI: https://doi.org/10.1145/3457607
[49] NASA. Astronauts. (????). Retrieved from https://github.com/DataResponsibly/Datasets.
[50] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. nyu Press.
[51] Department of Health and Human Services. Medical Expenditure Panel Survey. (????). Retrieved from https://meps.ahrq.gov/mepsweb/.
[52] Bureau of Transportation Statistics. National Summary of U.S. Flights. (????). Retrieved from https://www.transtats.bts.gov.
[53] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-mediated Communication* 12, 3 (2007), 801–823.
[54] Dana Pessach and Erez Shmueli. 2020. Algorithmic fairness. arXiv:2001.09784. Retrieved from https://arxiv.org/abs/2001.09784.
[55] ProPublica. Correctional Offender Management Profiling for Alternative Sanctions. (????). Retrieved from https://github.com/propublica/compas-analysis.
[56] John Rawls. 1971. *A Theory of Justice*. Harvard University Press.
[57] Richard V. Reeves and Dimitrios Halikias. 2017. Race gaps in SAT scores highlight inequality and hinder upward mobility. (2017). Retrieved from https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility.
[58] UCI Machine Learning Repository. German Credit. (????). Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/.
[59] John E. Roemer. 2002. Equality of opportunity: A progress report. *Social Choice and Welfare* 19, 2 (2002), 405–471.
[60] SAT. SAT. (????). Retrieved from https://www.qsleap.com/sat/resources/sat-2014-percentiles.
[61] Semantic Scholar. Semantic Scholar Open Research Corpus. (????). Retrieved from https://api.semanticscholar.org/corpus/.
[62] Law School Admission Council Research Report Series. LSAC National Longitudinal Bar Passage Study. (????). Retrieved from https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/LawStudents.
[63] Stephanie A. Shields. 2008. Gender: An intersectionality perspective. *Sex Roles* 59, 5–6 (2008), 301–311.
[64] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2219–2228.
[65] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. arXiv:1902.04056. Retrieved from https://arxiv.org/abs/1902.04056.
[66] Charles Spearman. 1961. The proof and measurement of association between two things. (1961).
[67] StackExchange. StackExchange. (????). Retrieved from https://stackexchange.com/.
[68] Julia Stoyanovich, Ke Yang, and H. V. Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the 21th International Conference on Extending Database Technology, Vienna, Austria, March 26–29, 2018*. 241–252. https://doi.org/10.5441/002/edbt.2018.22
[69] Engineering students. Engineering Students. (????). Retrieved from https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/EngineeringStudents.
[70] Supreme Court of the United States. 2009. Ricci v. DeStefano (Nos. 07-1428 and 08-328), 530 F. 3d 87, Reversed and Remanded. Retrieved from https://www.law.cornell.edu/supct/html/07-1428.ZO.html. (2009).
[71] Indian Institute Of Technology. IIT-JEE. (????). Retrieved from https://indiankanoon.org/doc/1955304/.
[72] The College Board. 2014. SAT Percentile Ranks. (2014).

[73] TREC. W3C Experts. (????). Retrieved from https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/TREC.
[74] Linda F. Wightman and Henry Ramsey. 1998. *LSAC National Longitudinal Bar Passage Study*. Law School Admission Council.
[75] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2536–2544. DOI: https://doi.org/10.1145/3219819.3220087
[76] Xing. XING. (????). Retrieved from https://github.com/MilkaLichtblau/xing_dataset.
[77] Yahoo. The Yahoo Webscope Program. (????). Retrieved from https://webscope.sandbox.yahoo.com/.
[78] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced ranking with diversity constraints. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 6035–6042. DOI: https://doi.org/10.24963/ijcai.2019/836
[79] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2021. Causal intersectionality and fair ranking. In *Proceedings of the Symposium on Foundations of Responsible Computing*. DOI: https://doi.org/10.4230/LIPIcs.FORC.2021.7
[80] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 22.
[81] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1569–1578.
[82] Meike Zehlike and Carlos Castillo. 2018. Reducing disparate exposure in ranking: A learning to rank approach. arXiv:1805.08716. Retrieved from https://arxiv.org/abs/1805.08716.
[83] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2017. Matching code and law: Achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery* 34, 1 (2017), 1–38.
[84] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*. http://jmlr.org/proceedings/papers/v28/zemel13.html.
[85] Yi Zhang. Yow News Recommendation. (????). Retrieved from https://www.younow.com.
[86] Yi Zhang. 2005. *Bayesian Graphical Model for Adaptive Information Filtering*. Ph.D. Dissertation. Carnegie Mellon University.

Received 8 June 2021; revised 17 April 2022; accepted 23 April 2022