Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems

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 subfields. 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 the first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this paper, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In this 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

1 INTRODUCTION

This is the second part of a survey on fairness in ranking. In the first part, we argued for the importance of a systematic overview of work on incorporating fairness requirements into algorithmic rankers. Which specific fairness requirements a decision maker will assert depends on the values and norms they are operationalizing, and on their bias mitigation objectives. To support an explicit mapping between these values, norms, and mitigation objectives, we introduced four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this paper. We recall a mind map that gives a visual overview of the frameworks in Figure 3.

In Part I of the survey, we described the preliminaries and fixed notations for score-based ranking, and then discussed existing technical methods for fairness in such rankings. We also discussed evaluation datasets that are used by the technical methods surveyed in this paper. In Part II of the survey we discuss technical work

Authors' addresses: Meike Zehlike, meikezehlike@mpi-sws.org, Humboldt University of Berlin, and Max Planck Institute for Software Systems, and and Zalando Research, Germany; Ke Yang, ky630@nyu.edu, New York University, NY, and University of Massachusetts, Amherst, MA, USA; Julia Stoyanovich, stoyanovich@nyu.edu, New York University, NY, USA.

0360-0300/2022/1-ART1 $15.00
https://doi.org/10.1145/3533380

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 $\mathcal{G}_2$

on fairness in supervised learning-to-rank and highlight representative examples of recent fairness methods in recommendation and matchmaking systems. To position Part II within the overall context of the survey, we start with a roadmap.

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 introduced 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 recommendation and matchmaking systems 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 from college admissions presented in Part I of the survey to discuss 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 Supervised learning to rank

In supervised learning to rank (LtR), we are given a set C of candidates; each candidate is described by a set $\mathbf{X}$ of features, including also sensitive features $\mathbf{A} \subseteq \mathbf{X}$. Each candidate $a \in C$ has an associated score attribute Y, which describes their quality with respect to a given task (e.g., college admissions). Every such association forms

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

(a)

$\boldsymbol{\tau}$
d
l

(b)

$\hat{\boldsymbol{\tau}}$
l
d

(c)

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. Scores Y_1 and Y_2 are the respective ground truth scores for queries $Q_1, Q_2 \in \mathbf{Q}$. We randomly distribute C into a training dataset $C_{train} = \{b, c, e, f, k, o\}$ and a test dataset $C_{test} = \{d, l\}$ (blue lines); (b) The ground truth ranking of C_{test} for query Q_1. The ranking model $\hat{f}$ should reproduce this ordering of candidates d and l when presented C_{test}; (c) A ranking predicted by a model with a bias against women. Note that by randomly choosing candidates d and l for C_{test} we accidentally injected a bias against women into our training data (all women are now ranked below men). A learning model is likely to pick up this bias and wrongly assign feature A_1 a high weight. (We remark that this is a very simple example on data bias for illustrative purposes).

an instance of either the training dataset C_{train} or the test dataset C_{test}. Like score-based rankers, LtR rankers compute candidate scores and return a ranking $\boldsymbol{\tau}$ with the highest-scoring candidates appearing closer to the top (per Eq.(1) in first part). The difference between score-based and LtR rankers is in how the score is obtained: in score-based ranking, a function is given to calculate the scores Y, while in supervised learning, the ranking function $\hat{f}$ is learned from a set of training examples and the score $\hat{Y}$ is estimated.[1]

Figure 2 describes the LtR process. We are given two datasets C_{train} and C_{test}.[2] We use C_{train} to train an LtR model, learning a ranking function $\hat{f}(\mathbf{X})$ that minimizes the prediction errors on Y_{train}. This is usually done by minimizing the sum of the individual errors $\hat{f}$ makes between the ground truth Y and its prediction $\hat{Y}$ for C_{train}. To evaluate the performance of the model $\hat{f}$, we apply it to C_{test}, and then compare ground truth scores and predictions. If model testing succeeds, meaning that the ranker's predictions are deemed sufficiently accurate, then $\hat{f}$ is deployed: a new set of candidates is ranked by predicting their scores $\hat{Y} = \hat{f}(\mathbf{X})$, and ranking the candidates according to these predictions.

As an example, consider Figure 1 that revisits our college admissions example from Figure 1 in the first part in a supervised learning setting. We are given six features, as previously described, and two ground truth scores Y_1 and Y_2 for each candidate.

First, training data is prepared as input: We divide the data into a training set $C_{train} = \{b, c, e, f, k, o\}$ and a test set $C_{test} = \{d, l\}$ (blue lines). Then a model is trained and tested using any available LtR method, such as RankNet [8] or ListNet [10]. Ranking $\boldsymbol{\tau}$ in Figure 1b depicts the ground truth ranking of C_{test} based on score Y_1.

[1]Note that the literature distinguishes point-wise, pair-wise and list-wise LtR methods and that Y has a slightly different meaning for each of them [29]. However, because the overall procedure remains the same, we will focus on point-wise LtR in the remainder of this section, and give technical details for pair-wise and list-wise methods in later sections, as appropriate.

[2]To follow machine learning best practices, we may also produce a separate validation dataset, used to tune model hyperparameters. We leave this out from our discussion for brevity.

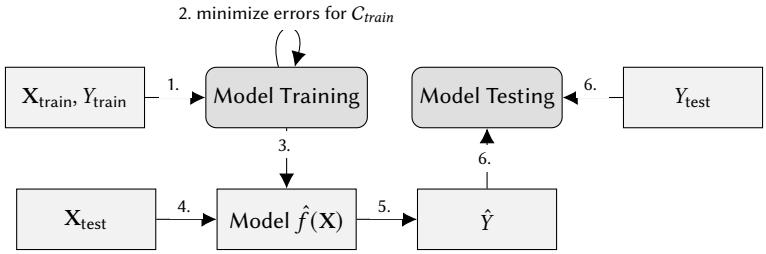

Fig. 2. Functional principle of supervised learning in ranking, commonly denoted learning-to-rank, or LtR. (1.) Training data C_{train}, consisting of tuples $(\mathbf{X}, Y)$, is given as input to an LtR algorithm that (2.) trains a ranking function $\hat{f}(\mathbf{X})$ (3.). This is done by minimizing the errors $\hat{f}(\mathbf{X})$ makes when predicting the scores $\hat{Y}$ for C_{train}. (4.) To test predictive accuracy of the model, the features $\mathbf{X}_{\text{test}}$ of C_{test} are given as input to $\hat{f}$ to (5.) predict scores $\hat{Y}$. (6.) Then, $\hat{Y}$ is compared to the ground truth Y_{test}.

Prediction accuracy. In traditional supervised learning, the term *utility* is often used to refer to prediction accuracy of $\hat{f}$. A common measure of prediction accuracy in LtR is the Normalized Discounted Cumulative Gain (NDCG) [22], which compares a ranking generated by model $\hat{f}$ to a ground-truth ranking (sometimes called the "ideal" ranking). NDCG measures the usefulness, or *gain*, of the candidates based both on their scores and on their positions in the ranking. NDCG incorporates position-based discounts, capturing the intuition that it is more important to retrieve high-quality (according to score) candidates at higher ranks, and is hence closely related to Eq. (3) in Section 2.1 of the first part. NDCG of a predicted ranking is computed relative to the gain of the ground-truth ranking, IDCG, and thus NDCG measures the extent to which the model is able to reproduce the ground-truth ranking from C_{train} in its predictions $\hat{Y}$. We are usually interested in NDCG at the top-k (denoted NDCG^k), and so normalize the position-discounted gain of the top-k in the predicted ranking by the position-discounted gain of the top-k in the ideal ranking (denoted IDCG^k), per Equation 1.

$$\text{NDCG}^k = \frac{1}{\text{IDCG}^k} \cdot \sum_{i=1}^{k} \frac{\hat{Y}\boldsymbol{\tau}(i)}{log_2(i+1)} \tag{1}$$

An important application of LtR are information retrieval systems, where users issue search queries, and expect the system to find relevant information and rank the results by decreasing relevance to their queries. Consider again our example in Figure 1, and suppose that we are additionally given a set $\mathbf{Q}$ of queries, each associated with C via a score.

In our example, two queries are given, Q_1 = "What are the most promising candidates to admit to a STEM major?" associated with score Y_1, and Q_2 = "What are the most promising candidates to admit to a humanities or arts major?" associated with Y_2. The training and test sets are formed by assigning the candidate features $\mathbf{X}$ and their respective scores Y per query: $C_{train} = \left\{(\mathbf{X}_{train}, Y_Q)\right\}_{Q \in \mathbf{Q}}$ and $C_{test} = \left\{(\mathbf{X}_{test}, Y_Q)\right\}_{Q \in \mathbf{Q}}$. With these sets as input, we can use the LtR procedure shown in Figure 2 to train a single model.

To evaluate model performance, its accuracy measures need to be extended to handle multiple queries. Commonly used measures are NDCG (Eq. 1) averaged over all queries and Mean Average Precision (MAP). MAP [33] consists of several parts: first, precision at position k ($P@k$) is calculated as the proportion of query-relevant candidates in the top-k positions of the predicted ranking $\hat{\boldsymbol{\tau}}$. This proportion is computed for all positions in $\hat{\boldsymbol{\tau}}$, and then averaged by the number of relevant candidates for a given query to compute average precision (AP). Finally, MAP is calculated as the mean of AP values across all queries. MAP enables a performance comparison between models irrespective of the number of queries that were given at training time.

Fairness. As is the case in score-based ranking, discussed in the first part of this survey, LtR methods may incorporate fairness objectives in addition to utility. Fairness interventions in LtR are warranted because the procedure described in Figure 2 is prone to pick up and amplify different types of bias (see Section 3.2 in the first part for details). For example, let us return to Figure 1 and note that, by randomly selecting $C_{train} = \{b, c, e, f, k, o\}$, in which all men are ranked above all women, we may have accidentally injected a strong gender bias into the learned model. This, in turn, may result in an estimated ranking $\hat{\boldsymbol{\tau}}$ in Figure 1c that places the male candidate l above the female candidate d, although their ground truth scores would place them in the opposite relative order. This may lead to biased future predictions for rankings that systematically disadvantage women ($\hat{\boldsymbol{\tau}}$ in Fig. 1c).[3]

As a remedy, two main lines of work on measuring fairness in rankings, and enacting fairness-enhancing interventions, have emerged over the past several years: probability-based and exposure-based. Both interpret fairness as a requirement to provide a predefined share of visibility for one or more protected groups throughout a ranking.

Probability-based fairness is defined by means of statistical significance tests that ask how likely it is that a given ranking was created by a fair process, such as by tossing a coin to decide whether to put a protected-group or a privileged-group candidate at position i [55, 57].

Exposure-based fairness is defined by quantifying the expected attention received by a candidate, or a group of candidates, typically by comparing their average *position bias* [23] to that of other candidates or groups.

$$\text{Exposure}\left(\boldsymbol{\tau}(i)\right) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi}\left[\mathbf{v}(\boldsymbol{\tau}(i))\right] \tag{2}$$

Here, $\pi : rnk(C) \rightarrow [0, 1]$ is the probability mass function over the ranking space, and position bias $\mathbf{v}(\boldsymbol{\tau}(i))$ refers to the observation that users of a ranking system tend to prefer candidates at higher positions, and that their attention decreases either geometrically or logarithmically with increasing rank [3, 23]. Logarithmic position-based discounting when computing exposure is in-line with position-based discounting of utility for score-based rankers (Eq. (3) in Section 2.1 of the first part) and with the NDCG measure for supervised LtR (Eq. 1).

The algorithmic fairness community is familiar with the distinction between individual fairness, a requirement that individuals who are similar with respect to a task are treated similarly by the algorithmic process, and group fairness, a requirement that outcomes of the algorithmic process be in some sense equalized across groups. Probability-based fairness definitions are designed to express strict group fairness goals. Thus, they do not allow later compensation for unfairness in higher ranking positions, since a ranking has to pass the statistical significance test at every position to be declared fair. If a ranking fails the fairness test at any point, it is immediately declared unfair, in contrast to exposure-based definitions. Exposure-based fairness can serve the goals of either individual fairness or group fairness, depending on the specific formalization. Individual unfairness in exposure can be expressed as the discrepancy $D(.)$ in exposure between two candidates a and b:

$$D(a, b) = \left|\text{Exposure}(a) - \text{Exposure}(b)\right| \tag{3}$$

Group unfairness can be expressed as the discrepancy $D(.)$ in the average exposure between two groups $\mathcal{G}_1$ and $\mathcal{G}_2$:

$$D(\mathcal{G}_1, \mathcal{G}_2) = \left|\frac{1}{|\mathcal{G}_1|}\sum_{a\in\mathcal{G}_1}\text{Exposure}(a) - \frac{1}{|\mathcal{G}_2|}\sum_{b\in\mathcal{G}_2}\text{Exposure}(b)\right| \tag{4}$$

Note that consensus on a definition of exposure has not yet been found and, while many measures feature position bias in some way, they disagree on its importance. An additional distinctive characteristic of fairness definitions is that some of them consider a notion of a candidate's merit when measuring disparities in exposure, while others explicitly leave it out. Most of the former understand merit as the utility score Y at face value. However, as we discussed in Section 3.3 of the first part, the understanding of merit depends on worldviews and

[3]We denote that this is a very simplified example for technical bias which we use for illustrative purposes.

on one's conception of equal opportunity. In Sections 3 and 4 we will present different interpretations of merit and exposure that have been used by LtR methods in information retrieval and recommender systems.

3 FAIR SUPERVISED LEARNING

In this section, we present several methods for fairness in learning-to-rank and information retrieval. We will continue to use the notation that we introduced in Section 2.1 wherever appropriate to illustrate the commonalities of the fields.

As we did in Part I of the survey, we will categorize the technical methods according to four normative frameworks, discussed in detail in Part I, Section 3. Figure 3 gives a structural overview of these frameworks and their sub-categories in the form of a *mind map*. For each method, we will highlight which normative choices they make by representing them on this mind map. Table 2 summarizes the categorization of all methods that are surveyed in the remainder of this document.

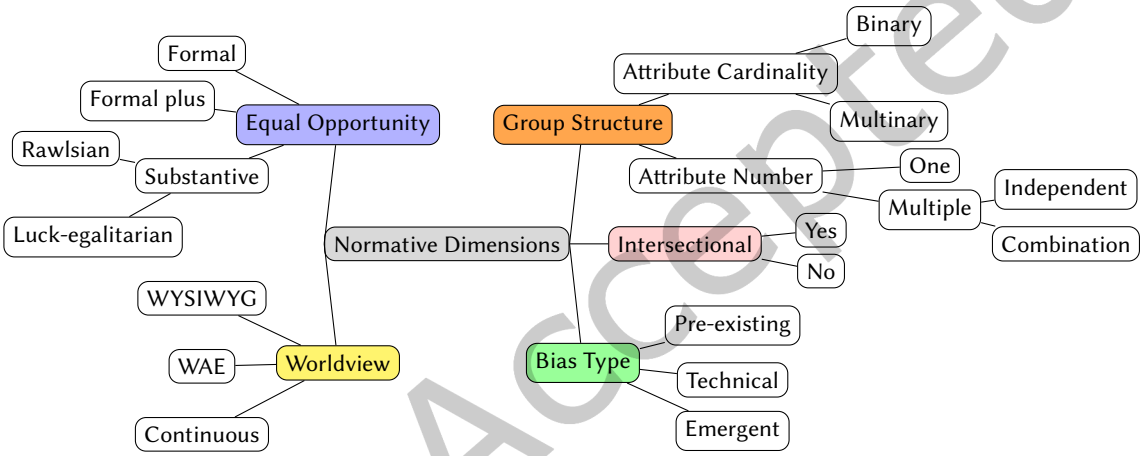

Fig. 3. A mind map summary of the structure of four classification frameworks. See Part I, Section 3, for details.

3.1 Pre-Processing Methods: Learning Fair Training Data

Pre-processing approaches are usually concerned with biases in the training data which they try to mitigate. Those biases can be of all three types: pre-existing biases appear in any data collection procedure in various ways. It is the way we interrogate, the decision which information we collect and which not, etc. Technical bias makes its way into data as rounding errors, different number and category encoding or the strategy choice how to handle missing values. Emergent bias arises when data is used in a different way than intended during collection. General advantages of pre-processing methods are:

- Pre-processing methods consider fairness as first concern in the machine learning pipeline.
- Most in- and post-processing methods rely on the availability of group labels during or after training, respectively. Pre-processing approaches instead commonly operate on a distance measure between individuals which allows to be agnostic to group membership. It is sufficient to define who should be similar to whom, based on the features that are available.
- Additionally it is possible to control for certain types of fairness across groups, even if only sparse information about group membership is available [27].

Table 2. Summary of method classification. "Pre-, in-" and "post-proc" refer to whether a method can be classified as pre-, in- or post-processing. "Binary" vs. "multinary" tells whether the method can handle two or more protected groups per one attribute at a time (e.g. young/old is a binary manifestation of the attribute "age", while kid/teen/adult/old is a multinary manifestation of said attribute). Since none of the methods in Part II of the survey handle intersectional discrimination, and we omit that column here. See Part I, Section 3 for a detailed description of the classification framework.

Method	Mitigation Point	Group structure	Bias	Worldview	EO Framework
iFair [26]	pre-proc.	multiple multinary attr.; independent	technical	WYSWYG	formal
DELTR [58]	in-proc.	one binary attr.	pre-existing	WAE	luck-egalitarian
Fair-PG-Rank [43]	in-proc	one binary attr.	technical	WYSIWYG	formal
Pairwise Ranking Fairness [4]	in-proc.	one binary attr.	?	WYSIWYG	formal-plus
FA*IR [57] & [60]	post-proc.	one multinary attr.; combination	pre-existing	continuous	formal / luck-egalitarian
Fair Ranking at LinkedIn [19]	post-proc.	one multinary attr.; combination	pre-existing; tec	continuous	none / luck-egalitarian (1 sensitive attr.)
CFAθ [59]	post-proc.	multiple binary attr.; combination	pre-existing	continuous	formal / substantive
Fairness of Exposure [42]	post-proc.	one binary attr.	pre-existing/ technical	WYSIWYG / WAE	formal / luck-egalitarian
Equity of Attention [6]	post-proc.	one multinary attr.; independent	technical / emergent	WYSIWYG	formal

candidate	A_1 (sex)	A_2 (race)	X_1 (GPA)	X_2 (SAT)
e	male	White	5	4
f	male	Asian	5	4
g	female	Black	5	3
h	female	White	5	3

Fig. 4. Dataset of college applicants. The goal is to find a fair feature representation and to ensure that a learning-to-rank method considers candidates based only on their non-sensitive features (here, X_1 and X_2). Assuming Manhattan distance as the distance measure d between two candidates, candidates e and f, as well as g and h should have the same fair feature representation $\widetilde{\mathbf{X}}$, where A_1 and A_2 are irrelevant for the ranked outcome.

General disadvantages are:

- Machine learning methods that rely on a separate feature engineering step are not applicable because the features identified by domain experts may be rendered meaningless, if fair representations are learned from the raw data.
- Current methods only operationalize individual fairness and treat group fairness as a special case of it.

3.1.1 iFair [26]. **Fairness Definition.** This work operates on an individual fairness objective to learn fair representations of training data points. It is based on the fairness definition by Dwork et al. [15], which states that similar individuals should be treated similarly. The goal is to transform an input record $\mathbf{X}_a$ (the feature vector for candidate a) into fairer data representations $\widetilde{\mathbf{X}}_a$ using a mapping ϕ, such that two individuals a and b, who are indistinguishable on their non-sensitive attributes $\mathbf{X} \setminus \mathbf{A}$ (marked by $\mathbf{X}^*$) should also be nearly indistinguishable in their fair representations $\phi(\mathbf{X}_a)$ (where sensitive attributes are included):

$$\left| d\left(\phi(\mathbf{X}_a), \phi(\mathbf{X}_b)\right) - d(\mathbf{X}_a^*, \mathbf{X}_b^*) \right| \leq \epsilon$$

Note that this definition assumes that a similarity measure d is available that can correctly (and free of bias) capture the differences between two individuals. As the paper uses the family of Minkowsky p-metrics, let us assume we choose the Minkowsky metric with $p = 1$ (i.e., the Manhattan distance as d). Reconsidering our college admission example, in Figure 4 we see that candidates e and f, as well as g and h have a distance of 0 in their non-protected features: $d(\mathbf{X}_e^*, \mathbf{X}_f^*) = d(\mathbf{X}_g^*, \mathbf{X}_h^*) = 0$. When comparing candidates across groups we see that the female group shows a Manhattan distance of 1 to the male group: $d(\mathbf{X}_e^*, \mathbf{X}_g^*) = d(\mathbf{X}_f^*, \mathbf{X}_g^*) = d(\mathbf{X}_e^*, \mathbf{X}_h^*) = d(\mathbf{X}_f^*, \mathbf{X}_h^*) = 1$. The proposed algorithm, *iFair*, would create a new feature set $\phi(\mathbf{X})$ that preserves those distances *and* includes sensitive attributes in such as way as to break correlations with the non-sensitive attributes. In our example, A_1 correlates with X_2, which may be picked up by a ranking model. To avoid this, iFair would assign non-correlating values to A_1, for example, by swapping the values of A_1 for candidates b and d.

Insights. Though not clearly stated, the wording suggests that attributes are measured in observable space *OS* and the definition seeks to reduce technical bias. Depending on the choice of distance metric d, the method would potentially be capable of learning representations that ignore *all* information about group membership, even if it is encoded partly in the non-protected features. In this case it would assign low weights to those non-protected features that indirectly encode protected ones, thus suggesting a leaning towards WAE. However, the choice of Minkowsky metrics, where distances are measured in terms of absolute numbers, suggests that the authors assume construct space $CS \sim OS$ and, hence, a leaning to a WYSIWYG worldview and to formal equality of opportunity. As stated above, the actual normative values that are incorporated into this method depend on the choice of d.

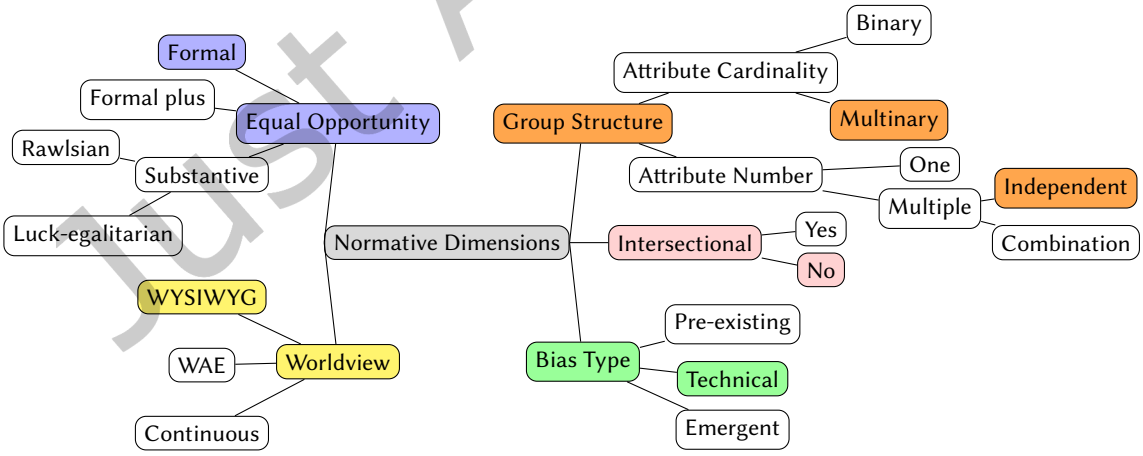

Fig. 5. Summary of the normative values encoded by iFair (Lahoti et al. [26]). This translates into normative choices that are implicitly taken when applying the method. Note, however, that this analysis is highly dependent on the choice of the metric d, which measures the similarity of two individuals.

Note also that the authors define fairness based on individuals, and group notions are not mentioned at all. While we can assume that individuals may have more than one protected attribute, it is not clear whether and how intersectional discrimination is a concern to the authors. Figure 5 and Table 2 summarize our analysis.

The authors do not address the question of whether Minkowsky metrics are prone to reproducing biases from training data, which may arise from biased observation processes. They tackle this problem in a follow-up work [27], where the distance metric is replaced by a *fairness graph* that captures pairwise similarities between individuals. In the graph, a node represents an individual a, and an edge between two nodes indicates that these individuals are to be considered similar. This approach has two advantages: it allows a comparison of an individual's non-protected attributes across different domains (e.g., the h-index of a successful researcher in programming languages is typically lower than that of a successful researcher in machine learning), and it allows for sparse similarity judgments, as individuals can be grouped into clusters based on their in-group relevance scores (e.g., the top-10% in the female group).

We do not further present Lahoti et al. [27] here, because the paper focuses on classification tasks in its experimental section. Pre-processing methods claim to be application-agnostic, however, to the best of our knowledge, Lahoti et al. [26] is the only work to-date that has been shown to work for ranking tasks.

Algorithm. The problem is formalized as a probabilistic clustering problem: given K clusters of similar individuals (with $K < n$), each is represented by a prototype vector v_k. A candidate record $\mathbf{X}_a$ is assigned to one of the v_k based on a record-specific probability distribution P_a that reflects the distances of the record from each of the prototype vectors, and thus forms the fair representation:

$$\phi(\mathbf{X}_a) = \widetilde{\mathbf{X}}_a = \sum_{k=1..K} P_{ak} \cdot v_k$$

This is used to formalize a utility objective that ensures a low reconstruction loss, and a fairness objective that demands that ϕ should preserve pair-wise distances on non-protected attributes between data records:

$$L_{\text{util}}(\mathbf{X}, \widetilde{\mathbf{X}}) = \sum_{a \in C} ||\mathbf{X}_a - \widetilde{\mathbf{X}}_a||_2$$

$$L_{\text{fair}}(\mathbf{X}, \widetilde{\mathbf{X}}) = \sum_{a,b \in C} \left(d(\widetilde{\mathbf{X}}_a, \widetilde{\mathbf{X}}_b) - d(\mathbf{X}^*_a, \mathbf{X}^*_b) \right)^2$$

The two objectives are combined into an objective function that the algorithm optimizes using gradient descent:

$$L = \lambda \cdot L_{\text{util}}(\mathbf{X}, \widetilde{\mathbf{X}}) + \mu \cdot L_{\text{fair}}(\mathbf{X}, \widetilde{\mathbf{X}})$$

The algorithm supports multiple groups.

Experiments. Experiments are performed on five real-world datasets and one synthetic dataset. Of these, XING [**?**] and AirBnB [**?**] are used for ranking tasks. Experiment on synthetic data show that representations learned by iFair remain nearly the same for all configurations, irrespective of changes in group membership. This means that changing the value of the sensitive attribute does not influence the learned representation, and so a model trained on such a representation will not learn any correlations between the sensitive attributes and other attributes.

The results show that applying learning algorithms on representations learned by iFair leads to more consistent decisions w.r.t. the distribution of items across a ranking than when applying the same algorithm to the original data. This means that two items with similar non-protected features will receive similar visibility in the resulting ranking.

3.2 In-Processing Methods: Learning a Fair Model

In-processing fair ranking methods extend the objective function of a learning-to-rank algorithm by a fairness term. Thus, the algorithm's optimization problem consists of an accuracy objective *and* a fairness objective, and the method learns to find the best balance between these two. General advantages are:

- In-processing methods yield better trade-offs between accuracy and fairness than post-processing methods, because finding this balance is at the heart of their learning objective [58].
- In-processing methods are capable of handling different types of underlying biases without knowing which particular type is present (see Section 3.2.1).

General disadvantages are:

- The impact of the fairness objective on the resulting ranking computed by an in-processing method is less apparent than for a post-processing method. The latter can make changes directly visible, while in the former case, separate models would have to be trained.
- Because of their goal to balance between fairness and accuracy, it is less clear what philosophical framework and worldview underlies fairness-aware models.

3.2.1 DELTR [58]. **Fairness Definition.** This method perceives unfairness as disparities in exposure, represented by the average visibility of a group (see Section 2.1 and Eq. 2). The exposure of a document is defined as its probability $P_{\hat{Y}Q}(\mathbf{X}_a)$ to appear in the top position of a ranking for query Q:

$$\text{Exposure}\left(\mathbf{X}_a|P_{\hat{Y}Q}\right) = P_{\hat{Y}Q}(\mathbf{X}_a) \cdot v_1 \tag{5}$$

where v_1 is the *position bias* of position 1, indicating its relative importance for users of a ranking system [22]. The exposure of group $\mathcal{G}$ is hence the average probability of its members to appear in the top position:

$$\text{Exposure}(\mathcal{G}|P_{\hat{Y}Q}) = \frac{1}{|\mathcal{G}|}\sum_{\mathbf{X}_a \in \mathcal{G}} \text{Exposure}\left(\mathbf{X}_a|P_{\hat{Y}Q}\right) \tag{6}$$

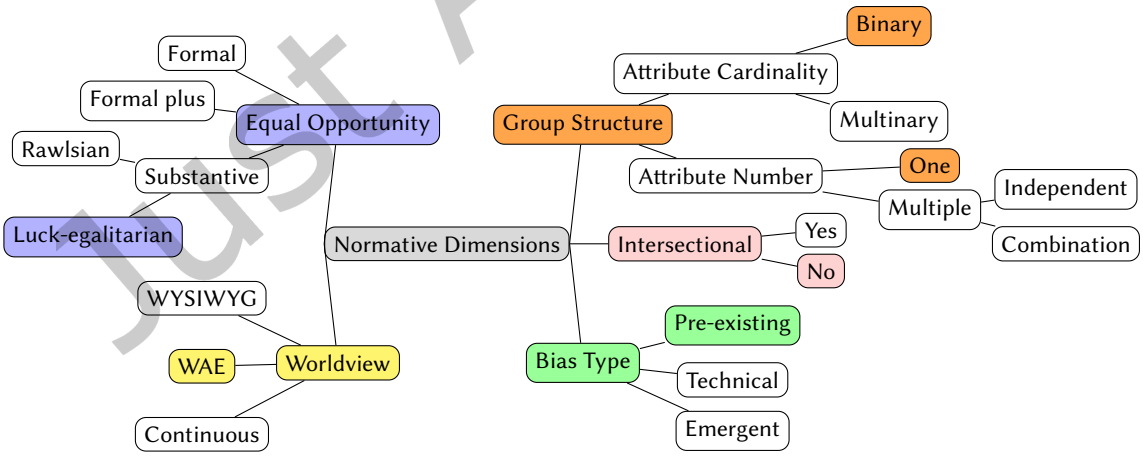

Fig. 6. Summary of the normative values encoded by DELTR (Zehlike and Castillo [58]). This translates into normative choices that are implicitly taken when applying the method. Note however, that because DELTR operates in a binary setting with only two groups, it can not handle multiple attributes.

Insights. As the definition optimizes for equality of exposure (exposure is the outcome), rather than equity, this means that the (potentially biased) qualification of a candidate is not taken into account, which suggests that its underlying assumption is a WAE worldview. We denote however, that because exposure is defined through the probability to appear in the top position, it indirectly contains a measurement of document relevance. We further denote that by setting γ to 0, the WYSIWYG worldview can also be adopted.

The method is concerned with pre-existing biases that lead to a biased observation process and, ultimately, to disparate exposure distributions. Under the assumption that the goal of the competition is to make future prospects comparable, this is consistent with luck-egalitarian EO. This method is agnostic to "true" score distributions and optimizes for equal exposure in decision space *DS* irrespective of whether score distributions in *CS* are different for demographic groups. For this reason, we argue that the mechanism is consistent with conditioning the qualification score on morally irrelevant characteristics (i.e., group membership), and place it into the category of luck-egalitarian EO. Figure 6 and Table 2 summarize our analysis.

Algorithm. The algorithm incorporates its unfairness measure into the objective function of a list-wise learning algorithm, namely ListNet [10], to simultaneously optimize for an accuracy metric L and an unfairness metric for two groups $D\left(\hat{Y}\right)$:

$$L_{\text{DELTR}}\left(Y, \hat{Y}\right) = L\left(Y, \hat{Y}\right) + \gamma D\left(\hat{Y}\right) \tag{7}$$

with

$$D\left(\hat{Y}\right) = \max\left(0, \text{Exposure}\left(\mathcal{G}_0 | P_{\hat{Y}Q}\right) - \text{Exposure}\left(\mathcal{G}_1 | P_{\hat{Y}Q}\right)\right)^2$$

where $\mathcal{G}_0$ denotes the non-protected group and $\mathcal{G}_1$ the protected one. The squared hinge loss is asymmetric to detect unfairness only if the protected group receives less exposure than the non-protected group, but not vice versa. The optimization problem is solved using gradient descent.

Experiments. To illustrate how the method works, let us return to our running example in Figure 1 on page 3: the algorithm gets as input the sensitive feature that forms a protected group, in this case A_1. As the training data C_{train} shows disparities in exposure for women (they are all ranked below the men), a standard learning-to-rank algorithm is likely to pick up A_1 as a predominant criterion for its model and assign a high weight to this attribute. DELTR instead will learn to ignore A_1 as a decision criterion because the unfairness metric penalizes ranking predictions that show high disparities in exposure for the different groups of A_1. In this way, DELTR can also compensate for systematic errors in the data or in the relevance measures (e.g., if the SAT test design favors male applicants, hence female applicants systematically receive lower scores).

Experiments are performed on three datasets, each exhibiting different types of bias, which are automatically handled by the proposed method, without explicit knowledge about what particular type of bias is present.

- **W3C experts:** [**?**] The experimental setup investigates situations in which bias is unrelated to relevance: expertise has been judged correctly, but ties have been broken in favor of the privileged group. In this case, including the sensitive feature during training yields very bad results in terms of disparate exposure and relevance, because all experts from the protected group are ranked at the bottom of the list.
- **Engineering students:** [**?**] The task is to predict a student's academic performance after the first year based on their admissions test results and school grades. Here, the same admissions test score relates to different levels of academic performance across groups: a score of 500 in the protected group relates to better academic performance than a score of 500 in the privileged group. This experiment investigates situations in which bias is coming from different score distributions among groups, and shows that a fair ranking can be derived without a degradation in accuracy. In this case, including the sensitive feature during training yields *better* results in terms of both exposure and relevance.

- **Law School Admission Council:** [?] These experiments show that disparities in exposure due to differences in academic performance can be reduced using a simple approach, but at a cost of accuracy. However, the trade-off is usually better (with lower accuracy loss) for in-processing methods than for post-processing methods.

3.2.2 Fair-PG-Rank [43]. **Fairness Definition.** The approach by Singh and Joachims [43] addresses technical bias that may be introduced by the ranking system itself due to *position bias* — giving candidates beyond the first few positions significantly less visibility compared to those who appear in the first positions. Like DELTR, Fair-PG-Rank operates based on a notion of document exposure. However, unlike DELTR, Fair-PG-Rank defines exposure as *expected attention*, which the authors consider to be equivalent to the expected position bias (as defined in Equation 2 in Section 2.1 on page 5). Similarly to Equity of Attention by Biega et al. [6], which we will discuss in Section 3.3.5, Fair-PG-Rank operates under a merit-based constraint: Each candidate $\boldsymbol{\tau}(i) = a$ in ranking $\boldsymbol{\tau}$ should receive exposure proportional to their utility $U(\boldsymbol{\tau}, a)$:

$$U(\boldsymbol{\tau}, a) \geq U(\boldsymbol{\tau}, b) \rightarrow \frac{\text{Exposure}(a)}{U(\boldsymbol{\tau}, a)} \leq \frac{\text{Exposure}(b)}{U(\boldsymbol{\tau}, b)} \tag{8}$$

This work further proposes a definition of individual fairness, per query Q, that measures the disparities in visibility $\mathbf{v}(.)$ for two candidates a, b in $\boldsymbol{\tau}$:

$$D(a, b|Q) = \frac{1}{|H^Q|} \sum_{(a,b) \in H^Q} \max \left[0, \frac{\mathbf{v}(a)}{U(\boldsymbol{\tau}, a)} - \frac{\mathbf{v}(b)}{U(\boldsymbol{\tau}, b)}\right] \tag{9}$$

with $H^Q = \{(a, b) \text{ s.t. } U(\boldsymbol{\tau}, a) \geq U(\boldsymbol{\tau}, b)\}$. The authors also propose a definition of group fairness for two groups, in which the notion of individual visibility of a document per Eq. 9 is replaced with group visibility:

$$D(\mathcal{G}_0, \mathcal{G}_1|Q) = \max \left[0, \frac{\mathbf{v}(\mathcal{G}_0)}{U(\boldsymbol{\tau}, \mathcal{G}_0)} - \frac{\mathbf{v}(\mathcal{G}_1)}{U(\boldsymbol{\tau}, \mathcal{G}_1)}\right] \tag{10}$$

with $\mathbf{v}(\mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{a \in \mathcal{G}} \mathbf{v}(a)$ being the average exposure of group $\mathcal{G}$.

Algorithm. Using the proposed fairness definitions, the authors extend the ListNet [10] ranking function to incorporate their disparity measures. The algorithm Fair-PG-Rank is learning an optimal ranking $\boldsymbol{\tau}^*$ via empirical risk minimization using the following learning objective:

$$\boldsymbol{\tau}^*_\delta = \arg\max_{\boldsymbol{\tau}} \frac{1}{N} \sum_{Q=1}^{N} [L(\boldsymbol{\tau})] - \lambda \frac{1}{N} \sum_{Q=1}^{N} [D(.|Q)] \tag{11}$$

with L being a loss function that measures the utility of $\boldsymbol{\tau}$ for the user. The optimization is done using gradient descent.

Insights. As already mentioned, the method is concerned explicitly with the inherent technical bias of a ranking that arises from showing ranked results in a one-dimensional list. It assumes a WYSIWYG world in which a document's merit in *OS* truly reflects its merit in *CS*, and no effort is made to reduce any errors that might have been introduced by the mapping function g. This method allocates outcomes (visibility) on the basis of merit (utility), and thereby codifies formal EO.

This method is explicitly concerned with equity of exposure. This means that documents being ranked will only receive as much exposure as they "deserve" based on their relevance. Returning to our college admission example (Figure 1, page 3), we see that in the training data (white background lines), all women have worse scores than men, as reflected in the relative ranking of the two groups. A standard LTR algorithm is therefore likely to treat A_1 as an important decision criterion and assign a high weight to it. Because Fair-PG-Rank optimizes for equity of exposure, this approach ensures that documents receive equal exposure among those that "deserve" it

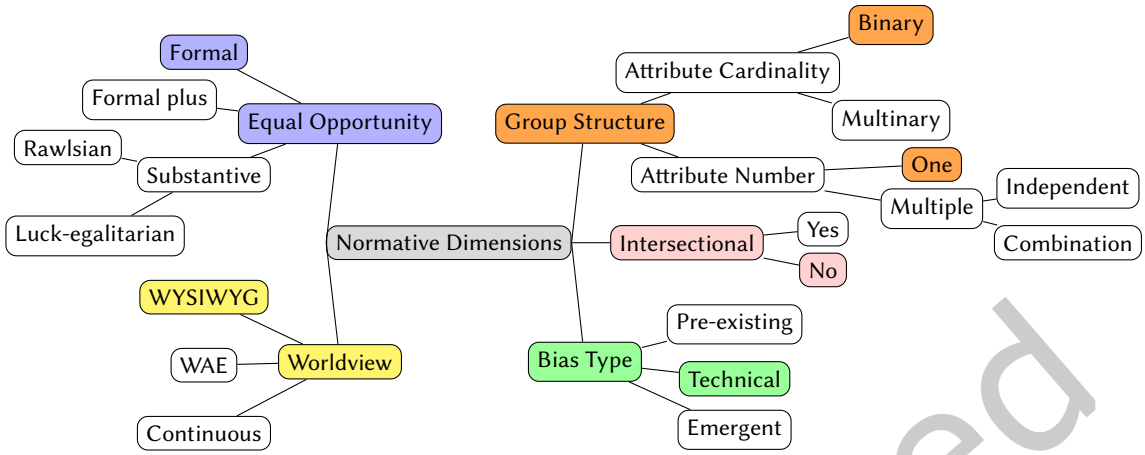

Fig. 7. Summary of the normative values encoded by Fair-PG-Rank (Singh and Joachims [43]). This translates into normative choices that are implicitly taken when applying the method.

based on their relevance. As such, Fair-PG-Rank ensures that a model is not discriminating based on a sensitive attribute, while assuming that relevance is computed correctly and does embed any patterns of discrimination. Note, that, similarly to DELTR, this method works for binary groups, and that it cannot handle multiple protected attributes. Figure 7 and Table 2 summarize our analysis.

Experiments. Experiments are done on a synthetic dataset, German credit [**?**] with the group fairness definition in Eq. 9 as part of the learning objective, and on the Yahoo! LTR dataset [**?**], with the individual fairness definition in Eq. 10 as part of the learning objective. The synthetic dataset has two features for each document, one of which is corrupted for the minority group G_1. The results show that, with increasing values of λ (see Eq. 11), the weight of the corrupted feature is decreasing. With the real world datasets, the authors show that the method works in a real world setting, but they do not discuss its implications. The authors compare their method to DELTR [58], but, because both their measure of disparate exposure and the type bias on which they are focusing are vastly different, it is not clear whether meaningful conclusions can be drawn from such a comparison. Further research is needed to understand whether and how methods with opposing worldviews, conflicting understandings of equality of opportunity, and different bias types can be compared.

3.2.3 Pairwise Fairness for Rankings, Beutel et al. [4]. **Fairness Definition.** This method is the first one to introduce a pairwise fairness metric for ranking predictions. The authors setup their method as a component of a cascading recommender system, however, it only considers the final ranking of items. We therefore present the method here, rather than in Section 4, where we talk about fairness in recommendation systems.

The framework is formalized as follows: Each query consists of user features $\mathbf{U}_i$ for user i and context features $\mathbf{C}$. Each ranking candidate a is described by a feature vector $\mathbf{X}_a$. Then, a ranker $\hat{f}$ is trained to predict user engagement, which relates to clicks $\hat{y}$ *and* interaction after a click $\hat{z}$ (such as purchase, ratings, etc.), which are then mapped to a scalar value to rank items: $\hat{f}(\mathbf{X}) = \hat{Y}$.

The focus of the fairness definition is on the risk for groups of ranked candidates to be under-recommended, under binary group membership (i.e., $A_a \in 0, 1$ for candidate a). For this, the authors first define pairwise accuracy, which describes the probability that a clicked candidate is ranked above another relevant unclicked candidate, for

the same query:

$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b\right)$$

Pairwise fairness asks if the pairwise accuracy is the same across the two groups:

$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 0\right) = P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 1\right)$$

To account for user engagement z, the definition is extended to compare only those candidates with each other that receive the same amount of engagement $\tilde{z}$:

$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 0, z_a = \tilde{z}\right) =$$
$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 1, z_a = \tilde{z}\right) \forall \tilde{z}$$

The authors further extend the definition to also consider group exposure in rankings, because two rankings could have the same pairwise accuracy across groups, while systematically putting candidates of one group to lower ranks of the list. To account for this, they split the definition into *intra-group* pairwise fairness:

$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = A_b = 0, z_a = \tilde{z}\right) =$$
$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = A_b = 1, z_a = \tilde{z}\right) \forall \tilde{z}$$

and *inter-group* pairwise fairness:

$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 0, A_b = 1, z_a = \tilde{z}\right) =$$
$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 1, A_b = 0, z_a = \tilde{z}\right) \forall \tilde{z}$$

Intra-group fairness indicates whether, across candidates from the same group, those that are more likely to be clicked are ranked above those that are less likely to be clicked. On the other hand, inter-group fairness describes whether mistakes of the ranker are at the cost of one particular group.

Insights. The framework as is is not clearly classifiable into WAE or WYSIWYG, because the authors talk about click probability and user engagement without further specifying what these two are composed off. Particularly they do not specify whether a candidate's merit is part of the click probability. Click through rate (CTR) is usually defined to contain some measure of relevance [38]. If that is the case here, then Y would contain a component that encodes merit, which is measured in *OS*, and hence the framework would correspond to WYSIWYG. The authors briefly mention that they assume the final ranking model $\hat{f}$ to only operate on relevant documents, which supports the assumption that the underlying worldview of this framework is WYSIWYG. We face the same uncertainty when thinking about which EO framework this work corresponds to. Without knowledge of the actual underlying estimation of click probability and user engagement, and without a statement on an individual's effort, it is not clear with which EO framework the fairness definition is consistent. However, because the authors state that they adopted a definition of "equal opportunity", and because their inter-group fairness objective balances error rates across groups, we map this method to formal-plus EO. An identification of the addressed bias is also not possible without the CTR definition.

Experiments. The experiments study the performance of the ranker with respect to the protected subgroup of candidates in the synthetic dataset, comparing the performance of this subgroup to the rest of the data, denoted by "not subgroup." (We will refer to these "not subgroup" candidates as "privileged" for ease of exposition.) The protected group represents approximately 0.2% of all items. The authors compare two versions of their approach: a model without any pairwise fairness constraint and one with an inter-group fairness constraint.

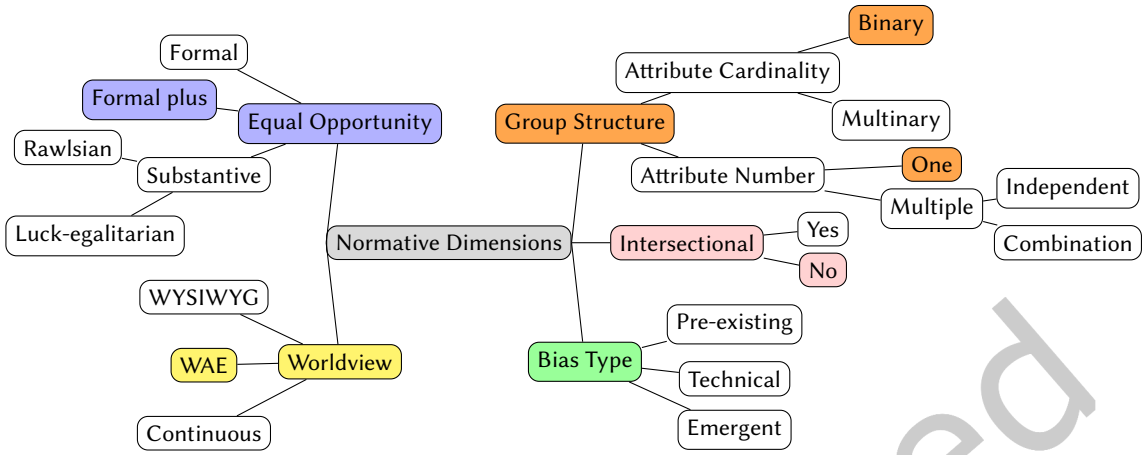

Fig. 8. Summary of the normative values encoded by Pairwise fariness for rankings (Beutel et al. [4]).

User engagement is grouped into four levels. Performance measures (pairwise accuracy) are aggregated across user engagement levels and averaged. Then, the ratio of the accuracy for the protected vs. privileged groups is computed, with a ratio of 1 corresponding to perfect fairness, a value above 1 corresponding to higher accuracy for the protected group, and a value below 1 corresponding to higher accuracy for the privileged group.

The overall pairwise fairness evaluation shows that the system under-ranks protected group candidates when the level of engagement is low, but, interestingly, it slightly over-ranks protected group candidates when the level of engagement is high. Intra-group pairwise fairness evaluation shows that, across all levels of engagement, the model has more difficulty selecting the clicked candidate when comparing protected group candidates than when comparing privileged group candidates. This is partly because the protected group is small and less diverse. Inter-group fairness evaluation shows that, across all levels of engagement, protected group candidates are significantly under-ranked relative to the privileged candidates. Further, the results show that the pairwise accuracy for the protected group in inter-group pairs is notably higher than in intra-group pairs, suggesting that protected group candidates, even when of interest to the user, are ranked below the privileged group candidates. When optimizing for inter-group fairness, these disparities are mitigated and protected group candidates receive more exposure (measure as the probability that candidate a is ranked above candidate b).

3.3 Post-Processing Methods: Re-Ordering Ranked Items

Post-processing algorithms assume that a ranking model has already been trained. A predicted ranking is handed to the algorithm, which re-orders items to improve fairness. Most algorithms operate on a notion of group membership, where certain groups are denoted as protected, while one group is denoted as non-protected (or privileged). As such, post-processing methods often model fairness constraints similarly to score-based fair ranking methods. We will reuse Figure 2 from Part I of the survey as our running example in this section, but will omit the non-sensitive features X_1, X_2 and X_3 because they are unimportant here.
General advantages of post-processing methods are:

- Many of them provide a guaranteed share of visibility for the protected group in the ranking.
- The effect the algorithms have on the ranked output is easy to visualize and understand, because the original ranking before the application of the fairness method can be compared to its result, in terms of how the items are re-ordered and, in some cases, in terms of the loss in a utility metric such as NDCG.

candidate	A_1 (sex)	A_2 (race)	$\hat{Y}$	v	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$
b	male	White	9	3.32	b	b	b	b
c	male	Asian	8	2.10	c	f	c	f
d	male	White	7	1.66	d	c	d	k
e	male	White	6	1.43	e	k	l	l
f	female	Asian	5	1.28	f	d	e	o
k	female	Black	4	1.18	k	l	f	d
l	female	Black	3	1.11	l	e	k	c
o	female	White	2	1.04	o	o	o	e

Fig. 9. A dataset C of college applicants. Score $\hat{Y}$ is predicted by a learning-to-rank model. v is the position bias of positions 1 – 8. Ranking $\hat{\tau}_1$ is predicted based on $\hat{Y}$, and the top-4 candidates will be admitted. Note that no female applicants are admitted in this scenario. A fair ranking $\hat{\tau}_2$ is produced according to the ranked group fairness condition of the FA*IR algorithm (Zehlike et al. [57]), with $p = 0.7$ and $\alpha = 0.1$. Ranking $\hat{\tau}_3$ is produced by the algorithm CFAθ with $\theta = 1$ if the dataset was changed to candidate f and k being male. Note that CFAθ cannot achieve a ranking where the only two female candidates l and o are both admitted. Ranking $\hat{\tau}_4$ with equal exposure across groups is produced by the method of Singh and Joachims [42].

p \ k	1	2	3	4	5	6	7	8	9	10	11	12
0.1	0	0	0	0	0	0	0	0	0	0	0	0
0.3	0	0	0	0	0	0	1	1	1	1	1	2
0.4	0	0	0	0	1	1	1	1	2	2	2	3
0.5	0	0	0	1	1	1	2	2	3	3	3	4
0.6	0	0	1	1	2	2	3	3	4	4	5	5
0.7	0	1	1	2	2	3	3	4	5	5	6	6

Table 3. Example values of the minimum number of protected items that must appear in the top-k positions, to pass the binomial ranked group fairness test with $\alpha = 0.1$. Table reproduced from Zehlike et al. [57].

General disadvantages are:

- The use of post-processing inherently suggests that fairness comes at the expense of accuracy, because the scores of a previously trained ranking model are taken as "the true anchor point." Depending on the properties of pre-existing bias in the training data, however, ranking models may incorporate biases that decrease accuracy as shown by Zehlike and Castillo [58], which renders any measurement of accuracy loss obsolete.
- Assuming biases are small, algorithms with a fixed fairness constraint [42, 57, 59] may cause a substantial loss in performance w.r.t. the original ranking. This may be the case if the score distribution of the protected group is very different than that of the privileged group, leading to systematically lower scores for protected group members.

*3.3.1 FA*IR, Zehlike et al. [57, 60].* **Fairness Definition.** This work builds on Yang and Stoyanovich [55], discussed in Part I of the survey, and adopts a fairness definition based on the assumption that rankings are fair when the decisions on candidate placement are drawn from a Bernoulli distribution (coin tosses) that is not impacted by the candidate's sensitive attributes. FA*IR [57] ensures that the number of protected candidates does not fall far below a required minimum percentage p at any point in the ranking, by formulating this fairness as a statistical significance test of whether a ranking was likely to have been produced by a Bernoulli process. In Zehlike et al. [60], the authors extend the mathematical framework of [57] to a multinomial distribution (roll of a dice), and provide a vector of minimum proportions $p_{\{\mathcal{G}\}}$, containing one p for each group $\mathcal{G}$. This way the extended algorithm can handle more than two groups at the same time, while the original operates in a binary group setting. Both methods are concerned with disparate impact, as they do not take any notion of merit into account for their re-ranking strategy. A ranking prefix of length k is considered to fairly represent the protected group if the proportion of protected group members α does not fall below the minimum target proportion $p_{\{\mathcal{G}\}}$:

$F\left(\boldsymbol{\tau}_{\{\mathcal{G}\}}\ k, p_{\{\mathcal{G}\}}\right)$, with F corresponding to the multinomial cumulative distribution function. If this condition holds for each prefix $k = 1, \ldots, n$, then the entire ranking is considered to be fair.

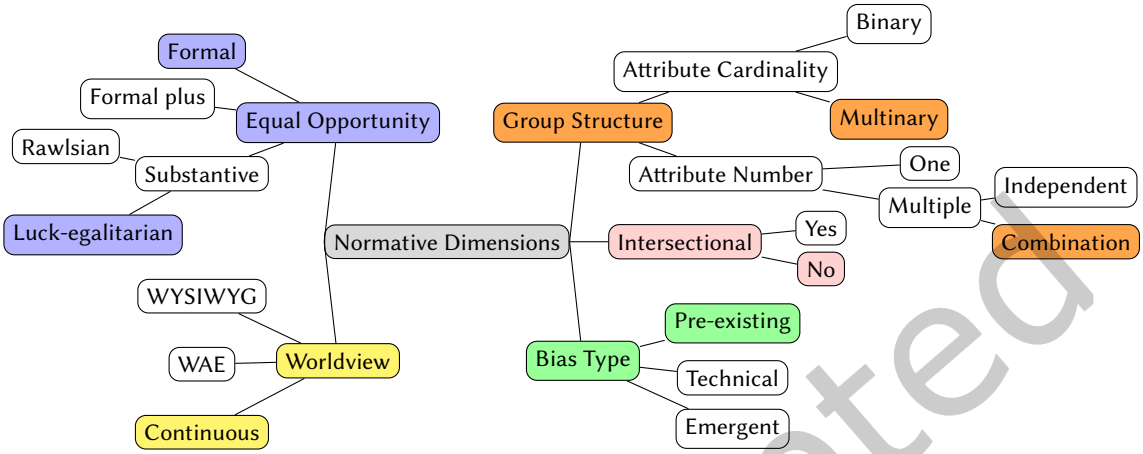

Fig. 10. Summary of the normative values encoded by FA*IR (Zehlike et al. [57, 60]).

Insights. The existence of minimum proportions for a protected group suggests a WAE worldview, and a focus on pre-existing bias. However, the WYSIWYG worldview can also be adopted for a particular group, if needed: the respective p can be set to a low value, meaning that only very few candidates from that group need to be in the ranking. This permits a gradual transition from WYSIWYG to WAE. However, this is a mere side effect of the method, and, in contrast to Zehlike et al. [59], this is not to be seen as a desideratum of FA*IR. Critically, if p is chosen too high, the framework will rank a lot more protected candidates in the highest positions, than non-protected ones. This would be justified only if one assumes *either* that the protected group is actually a lot more relevant (their scores are very high in the construct space CS), but the measurements in observable space are extremely biased; *or* that the scoring model produces inverted predictions for the protected group in decision space DS. While such cases do exist, it is questionable whether one should attempt to correct such a flawed ranking, rather than disregarding it altogether.

FA*IR assumes that differences between groups in the relevance distribution are an artifact of different circumstances for each group. Under the assumption that the goal of the fairness-enhancing intervention is to equalize access to opportunity over a lifetime, this method is consistent with substantive EO. Furthermore, because the proposed method conditions on morally irrelevant characteristics, we classify it as luck-egalitarian EO. Figure 10 and Table 2 summarize the mapping of the methods in this section to normative frameworks.

Algorithm. For performance improvements, the binomial algorithm [57] pre-computes a table that contains the minimum number of protected candidates from each group at each position in the ranking given a minimum proportion p (see Table 3). This is done by computing $F^{-1}(\alpha; k, p)$, the percent point function of the binomial CDF. Then, the two groups are ranked separately by decreasing scores, and are then merged into a single ranking according to the table. Whenever the ranking by score violates the minimum number requirement, a protected item is put at the respective position. For multiple protected groups the algorithm receives a minimum proportion for each group, and the data structure to hold the minimum numbers of candidates of each position becomes a tree instead of a table. Thus multiple ways are possible to order candidates to produce a ranking that is fair according to a multinomial distribution process.

As an example, consider Figure 9 where a model has predicted relevance scores $\hat{Y}$, based on which ranking $\hat{\tau}_1$ in Figure 9 is produced. This ranking is fair according to the binomial FA*IR algorithm, if the input is $p < 0.5, \alpha = 0.1$. If $p \geq 0.5$ the candidates have to be reordered. Ranking $\hat{\tau}_2$ in Figure 9 is produced by FA*IR with $p = 0.7, \alpha = 0.1$.

Experiments. The experimental evaluation in [57] is done on the three real datasets, COMPAS [**?**], German credit [**?**], and SAT [**?**], and shows the effects of the methods on performance in terms of NDCG and maximum rank loss. In [60], experiments are performed on COMPAS, German Credit, and LSAC [52]. All scenarios under consideration relate to questions of distributive fairness, where certain benefits (scholarship, visibility, pardon) are to be distributed fairly across two groups.

3.3.2 Fairness-Aware Ranking at LinkedIn [19]. **Fairness Definition and Algorithm.** In this work, fairness is defined through the minimum and the maximum number of candidates at each position in the ranking. A proportion $p_{\mathcal{G}}$ for each group $\mathcal{G}$ is given as input, and a ranking $\boldsymbol{\tau}$ is declared as fair for all groups if the following conditions hold:

$$\forall k \leq |\boldsymbol{\tau}|,\ \forall \mathcal{G} : \sum_{i=0}^{k} \boldsymbol{\tau}(i) \in \mathcal{G} \leq \lceil p_{\mathcal{G}} \cdot k \rceil$$

$$\forall k \leq |\boldsymbol{\tau}|,\ \forall \mathcal{G} : \sum_{i=0}^{k} \boldsymbol{\tau}(i) \in \mathcal{G} \geq \lfloor p_{\mathcal{G}} \cdot k \rfloor$$

To create a fair ranking, the algorithm first checks whether any group has not yet met its minimum number requirement (first inequality), and , if so, adds protected candidate to the ranking. If more than one group does not yet meet the requirement, the candidate with the highest utility score $\hat{Y}$ among all eligible ones is chosen. If all minimum requirements are met, the algorithm takes the highest-scoring candidate among all that have not yet exceeded their maximum count (second inequality). As soon as more than three groups are present, this can easily lead to an infeasible state, where the requirements can no longer be met, no matter which candidate is chosen. The authors therefore propose three additional ranking algorithms, each of them improving the likelihood to create a fair ranking by changing the strategy to choose the next candidate.

Insights. The algorithm can handle multinary protected attributes, and is concerned with both pre-existing and technical bias. As in FA*IR, the proposed methods operate on minimum proportions for protected groups, and the fairness constraints are not incorporating a notion of merit. The values of $p_{\mathcal{G}}$ can also be chosen freely and, therefore, at first glance, a spectrum of worldviews and EO frameworks can be adopted. However, there is an important detail in the algorithm that makes it inconsistent with WAE and with substantive EO: From all possible candidates to be selected, the algorithm always chooses the highest-scoring one. This is essentially the same design choice as was made by Celis et al. [11] in their Constrained ranking maximization paper (discussed in Section 5.2.1 of Part I of the survey) when the number of sensitive attributes exceeds 1.

This choice leads to an implicit integration of merit into the fairness objective, and it assumes that scores are comparable across groups. For this reason, we classify this method as being compatible with substantive EO, and, specifically, with the luck-egalitarian EO doctrine, for *one* sensitive group, but deem it inconsistent with any EO framework for multiple sensitive groups. It also ignores the fact that groups facing intersectional discrimination commonly show larger group skews and biases when their true merit from *OS* is translated into scores in *DS*. Figure 11 and Table 2 summarize our analysis.

Experiments. Experiments are performed on synthetic datasets with different numbers of sensitive attributes. The evaluation measures report fairness in terms of normalized discounted KL divergence [55], and performance in terms of NDCG [22]. Additionally, A/B-tests are performed after implementing the method as part of the

LinkedIn Recruiter product. The minimum proportions are set to match the distribution of all genders among the relevant candidates. The results show significant increases in terms of representativeness of gender among the top-k, but no decreases in performance measures, which is why the algorithm was implemented permanently in the product.

3.3.3 Continuous Fairness with Optimal Transport, Zehlike et al. [59]. **Fairness Definition and Algorithm.** This work defines a mathematical framework, CFAθ, to continuously interpolate between the WYSIWYG and WAE worldviews. The authors argue that, legally speaking, WYSIWYG is consistent with individual fairness, while WAE is consistent with current anti-discrimination law that defines group fairness in terms of *statistical parity of outcomes*. The answer to the question of what a fair distribution of outcome is, depends on the estimated extend of indirect discrimination and pre-existing bias in the scoring model. Interestingly, this means that any fairness definitions departing from group fairness as statistical parity, and individual fairness as meritocratic scores, do not yet have any legal meaning. Furthermore, the authors state that the current rulings on anti-discrimination cases only involve actual unfair decisions, and not "softer" disadvantages such as reduced visibility in a ranking.

As we discussion in Part I of the survey, generally speaking, the WYSIWYG worldview corresponds to the meritocratic ideal and, hence, is consistent with formal EO, while WAE corresponds to substantive EO. This is because an individual's measurable effort (here, their raw score) is seen to be drawn from different distributions μ_k per group in OS, while in CS, there exists only one ν, meaning that all groups have essentially the same distribution of true effort.

The framework can handle multiple sensitive attributes, each defining a partitioning over all individuals: $\mathbf{X} = \bigcup_{k \in \{0,1\}^N} g^{-1}(k)$, where $g : \mathbf{X} \to \{0,1\}^N$ is a mapping that returns 1 if an individual carries a certain trait from the set of N features. The k-th group is therefore $\mathcal{G}_k := g^{-1}(k)$. The authors explicitly include all features in the group definition instead of only taking certain attributes that are legally protected into account. This broad definition has the advantage that it can handle non-sensitive features that serve as proxies for the sensitive features.

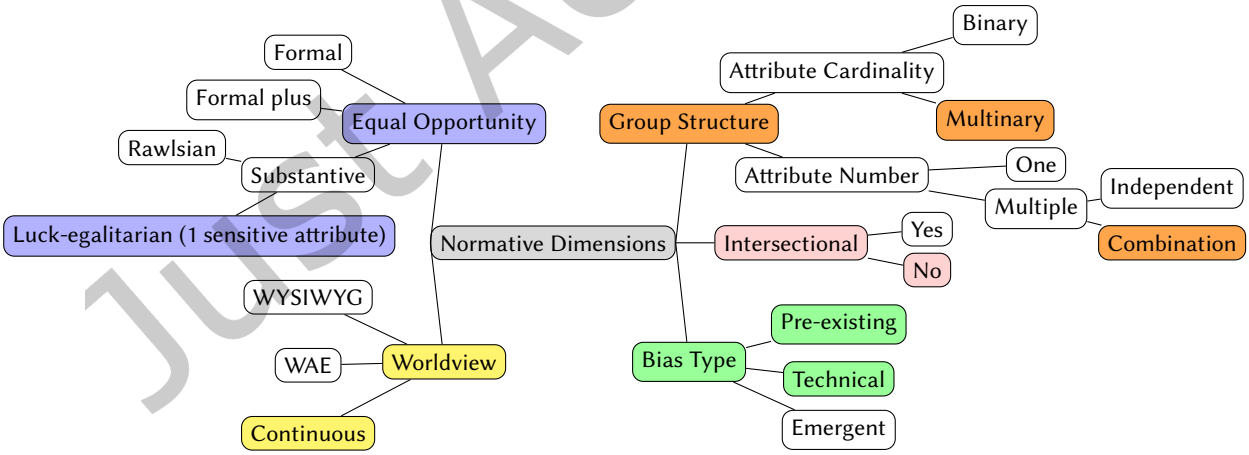

Fig. 11. Summary of the normative values encoded by the fair LinkedIn method (Geyik et al. [19]). This translates into normative choices that are implicitly taken when applying the method. Note, that in contrast to FA*IR, this method is not compliant with substantive EO for more than 1 sensitive attribute, even though the methods appear to be very similar. This shows how a small technical detail can lead to profound differences in underlying values.

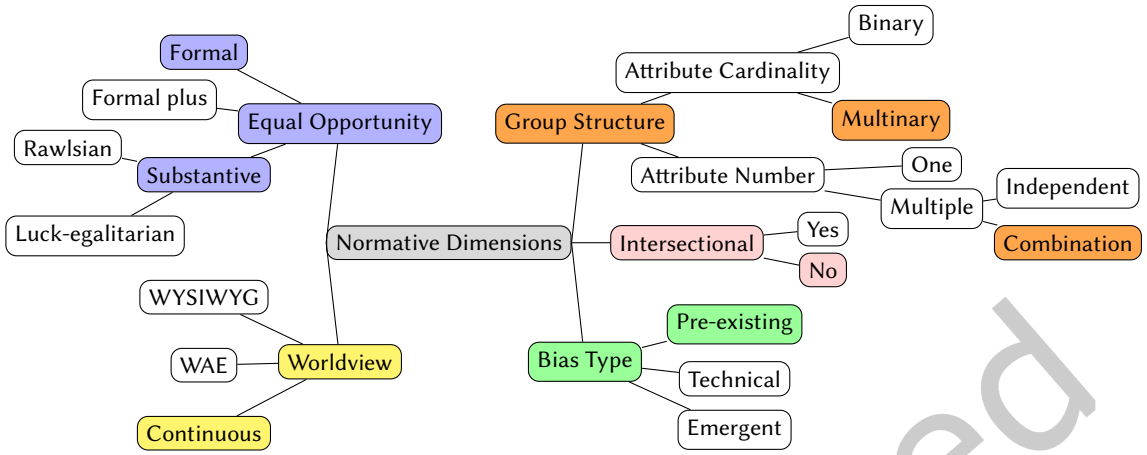

Fig. 12. Summary of the normative values encoded by CFAθ (Zehlike et al. [59]). The paper is a rare example of explicit statements on the incorporated values.

The framework assumes that a potentially biased scoring function S is given, and maps from the space of individual traits $\mathbf{X}$ to an n-dimensional vector $S : \mathbf{X} \rightarrow \mathbb{R}^n$, and that each group's score forms a probability distribution μ_k. The combined score distribution is called μ and corresponds to a metric from OS, which, as usual, is prone to pre-existing bias and other systematic errors with different group skews.

The authors then define a score distribution $\nu_k = \mu_k \circ T_k^{-1}$ to be the fair score representation of group k obtained by an optimal transport map $T_k : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Depending on the worldview, ν is defined differently: In WYSIWYG $\mu = \nu_k$ for all k groups, while in WAE any differences between the μ_k are solely the product of bias, and, there, there exists a single ν that is the same for all groups. In the former case, the optimal transport matrix is the identity matrix. In the latter, the WAE fair representation distribution ν that satisfies statistical parity is defined as the barycenter in Wasserstein space of the μ_k, and a T_k has to be found for each group to transform μ_k into ν while minimizing violations against decision maker utility and individual fairness. The framework further defines a displacement interpolation with a fairness parameter $\theta \in [0, 1]$, which allows to transform μ_k into any distribution $\mu_k^\theta = \mu_k \circ (T^\theta)^{-1}$ between the WYSIWYG (or individual fairness) policy μ and the WAE (or group fairness) policy ν. This means that a high θ corresponds to more emphasis on group fairness, while a low θ corresponds to more individual fairness, with $\mu^0 = \mu$ and $\mu^1 = \nu$.

A notable advantage of this approach is that it does not rely on the existence of a distance metric between individuals, in contrast to many other methods [15, 26]. This is important because it is not clear how such a distance metric, if it were actually available in OS, would be less prone to biases and errors than any other commonly used optimization metric.

Insights. As CFAθ moves the group distributions of predicted scores closer to each other, this means that when setting $\theta = 1$ the algorithm achieves statistical parity for each group throughout the ranking, *but not more so* as compared to FA*IR, which can push candidates from the protected group even higher. Such a setting would result in the same ranking $\hat{\boldsymbol{\tau}}_2$ as shown in Figure 9, given that there are 50% of male and female candidates in the dataset. However if there where only 25% female candidates in the dataset (e.g., if candidates f and k were male), then the algorithm would produce the ranking $\hat{\boldsymbol{\tau}}_3$ as shown in Figure 9 when given $\theta = 1$ as input. Note that with such a dataset the method cannot return a ranking in which the now only two female candidates are ranked among those admitted (i.e., in the top-4). Figure 12 and Table 2 summarize our analysis.

Experiments. The experimental evaluation is done on a synthetic dataset with 100,000 data points, a score feature and two sensitive features. Group membership for this experiment is defined by all combinations of the values of the sensitive features. Rankings are produced based on the score column and the performance of a fair ranking with different θ is measured in terms of NDCG. The fairness of a ranking is measured as the share of each group at each n positions, with n ranging from 10 to 1000.

A second experiment is conducted on the LSAC dataset [**?**]. The experiments confirm the general properties of post-processing methods: it is clearly visible how groups are distributed more evenly across all positions with increasing values of θ. However, depending on the differences between the μ_k, a processed ranking based on the fair representation v can show significant declines in performance measures w.r.t. the raw score ranking.

3.3.4 Fairness of Exposure, Singh and Joachims [42]. **Fairness Definition.** The fairness objective of this work is set as a linear combination $\mathbf{a}^T \mathbf{P}^{\boldsymbol{\tau}}_{a,i} \mathbf{v} = h$ with $\mathbf{a}$ being a vector to encode group membership, $\mathbf{P}^{\boldsymbol{\tau}}_{a,i}$ as the probability that $\hat{f}$ places candidate a at rank i in $\boldsymbol{\tau}$, and $\mathbf{v}$ reflecting the importance of a position in a ranking. This equation is solved under three different group fairness constraints based on a definition of exposure that a candidate a receives under $\mathbf{P}^{\boldsymbol{\tau}}$:

$$\text{Exposure}(\mathbf{X}_a|\mathbf{P}^{\boldsymbol{\tau}}) = \sum_{i=1}^{k} \mathbf{P}^{\boldsymbol{\tau}}_{a,i} \mathbf{v}(i)$$

with $\mathbf{v}(i)$ being the position bias of position i in ranking $\boldsymbol{\tau}$. The average exposure of a group $\mathcal{G}$ is defined as follows:

$$\text{Exposure}(\mathcal{G}|\mathbf{P}^{\boldsymbol{\tau}}) = \frac{1}{|\mathcal{G}|} \sum_{\mathbf{X}_a \in \mathcal{G}} \text{Exposure}(\mathbf{X}_a|\mathbf{P}^{\boldsymbol{\tau}})$$

The goal is to distribute exposure fairly between groups $\mathcal{G}_0$ and $\mathcal{G}_1$ using the following three definitions:

(1) **Demographic Parity** states that the average exposure of groups shall be equal $\text{Exposure}(\mathcal{G}_0|\mathbf{P}^{\boldsymbol{\tau}}) = \text{Exposure}(\mathcal{G}_1|\mathbf{P}^{\boldsymbol{\tau}})$.

(2) **Disparate Treatment** requires equity of exposure across groups (i.e., the average exposure in relation to their average utility should be equal across groups):

$$\frac{\text{Exposure}(\mathcal{G}_0|\mathbf{P}^{\boldsymbol{\tau}})}{U(\boldsymbol{\tau}, \mathcal{G}_0)} = \frac{\text{Exposure}(\mathcal{G}_1|\mathbf{P}^{\boldsymbol{\tau}})}{U(\boldsymbol{\tau}, \mathcal{G}_1)} \tag{12}$$

(3) **Disparate Impact** is measured in terms of disparate click through rates (CTR) [38] across group. The goal is to equalize CTR across groups, given the groups average utility:

$$\text{CTR}(\mathcal{G}|\mathbf{P}^{\boldsymbol{\tau}}) = \frac{1}{|\mathcal{G}|} \sum_{a \in \mathcal{G}} \sum_{i=1}^{k} \mathbf{P}^{\boldsymbol{\tau}}_{a,i} Y_a \mathbf{v}(i)$$

with Y_a being the relevance of candidate a.

$$\frac{\text{CTR}(\mathcal{G}_0|\mathbf{P}^{\boldsymbol{\tau}})}{U(\boldsymbol{\tau}, \mathcal{G}_0)} = \frac{\text{CTR}(\mathcal{G}_1|\mathbf{P}^{\boldsymbol{\tau}})}{U(\boldsymbol{\tau}, \mathcal{G}_1)} \tag{13}$$

Insights. The definition of demographic parity addresses the problem of pre-existing bias and corresponds to the WAE framework, as it tries to balance visibility across groups independently of their performance in *OS* (note, however, that a group's exposure depends on $\mathbf{P}^{\boldsymbol{\tau}}$, and may thus indirectly depend on a utility measure, if $\mathbf{P}^{\boldsymbol{\tau}}$ is calculated based on document utility). The definition assumes that $CS \nsim OS$ and therefore that an individuals true effort is different from the measured effort, accounted for by the demographic parity definition. Under the assumption that the goal of the competition is to equalize opportunity over a lifetime, these methods are

consistent with substantive EO. Further, because of conditioning on group membership, we map these methods to luck-egalitarian EO. Figure 13 and Table 2 summarize our analysis.

The definition of disparate treatment explicitly addresses the technical bias of a ranking, also known as position bias, by ensuring that all documents of the same utility receive equal visibility. This is consistent with formal EO framework. The method corresponds to the WYSIWYG worldview because document utility is measured through features from observable space without taking into account that a biased observation process may exist, hence $CS \sim OS$. Note that this does not necessarily correspond to individual fairness, because utility and exposure are averaged across individuals of a group, and can lead to a downgrading of high scoring individuals in otherwise badly performing groups. Figure 14 and Table 2 summarize our analysis.

The definition of disparate impact is misleading does not comply with the *legal* definition of disparate impact, which is described solely in terms of the deviation from statistical parity. Referring to this definition as "disparate impact" is misleading, because the click through rate contains a notion of document relevance. Statistical parity, in contrast, does not consider any relevance measure whatsoever, precisely because it assumes that these very measurements are subject to pre-existing biases and a biased mapping from CS to OS. As the given definition mostly corresponds to formal EO and a WYSIWYG worldview, it would be more appropriate to label it as a different version of disparate treatment that is concerned with click through rate instead of exposure. Since the two definitions correspond to the same value frameworks, we have summarized both of them in Figure 14, as well as in Table 2.

Algorithm. The algorithmic framework is implemented as an ILP that maximizes ranking utility given one of the above constraints translated into a scalar h. Note that this is in contrast to Biega et al. [6], to be discussed in Section 3.3.5, who instead constraint quality and optimize for disparate treatment:

$$\begin{aligned} \arg\max_{\mathbf{P}} \quad & Y^T \mathbf{P} \mathbf{v} \\ \text{subject to} \quad & \mathbb{1}^T \mathbf{P} = \mathbb{1}^T, \\ & \mathbf{P}\mathbb{1} = \mathbb{1}, \\ & 0 \leq \mathbf{P}_{a,i} \leq 1, \\ & \mathbf{a}^T \mathbf{P} \mathbf{v} = h \end{aligned} \tag{14}$$

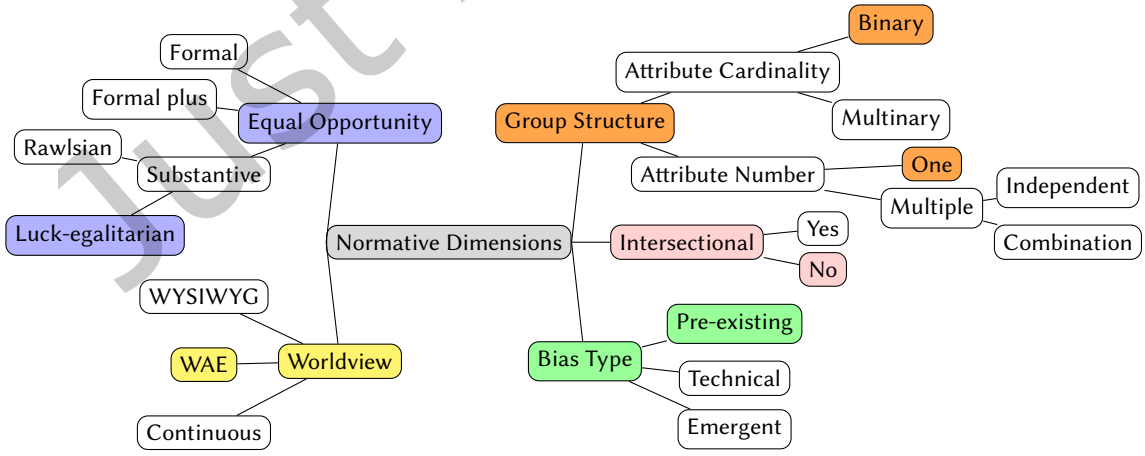

Fig. 13. Summary of the normative values encoded by the definition of demographic parity of Singh and Joachims [42].

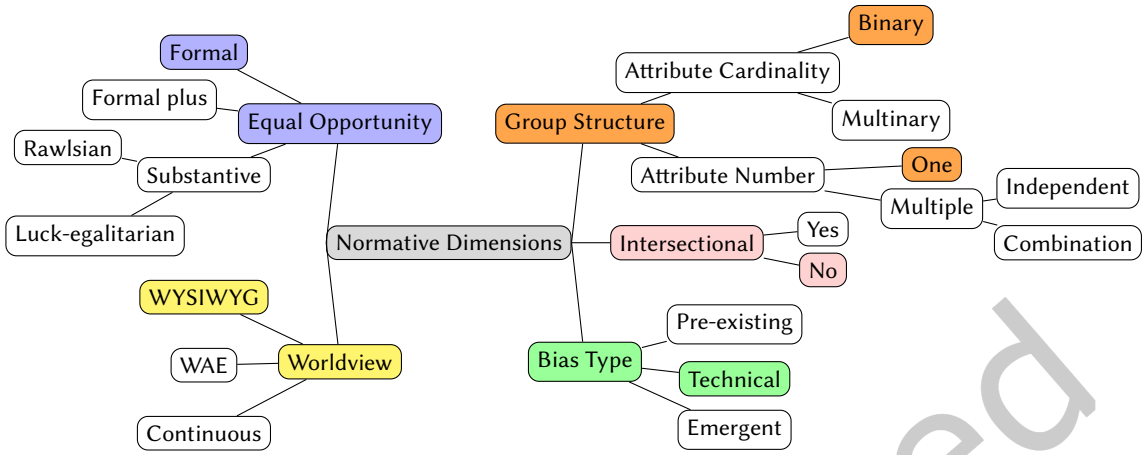

Fig. 14. Summary of the normative values encoded by the definitions of "disparate treatment" and "disparate impact" of Singh and Joachims [42].

Depending on the respective definition of fairness, the outcome rankings will look quite differently. Let us assume that the model is absolutely sure about where to place each candidate, such that $\mathbf{P}$ becomes the unit matrix and exposure of a group is calculated as the sum of each group member's position bias in the ranking. This gives us a group exposure of 2.13 for the male group and 1.15 for the female group for the ranking $\hat{\boldsymbol{\tau}}_1$ in Figure 9. Let us consider the demographic parity objective, that requires to equalize exposure for both groups. As the position bias $\mathbf{v}$ is a constant value, the algorithm will modify the scores $\hat{Y}$ until the parity objective is met. A possible solution is ranking $\hat{\boldsymbol{\tau}}_4$, shown in Figure 9, with group exposure of 1.66 for the male group and 1.62 for the female group.

The other two objectives work in an similar manner, except that they take utility of a ranking into account.

Experiments. The experiments are framed within three different scenarios of unfairness: biased allocation of opportunity (in job candidate rankings), misrepresentation of real-world distributions (biased Google image search for CEO), and fairness as freedom of speech (equality of voice within news media channels like YouTube or Twitter). For these, the authors create a synthetic dataset with 100,000 entries and a binary protected attribute. Furthermore they use the real YOW news recommendation dataset [**?**].

3.3.5 Equity of Attention, Biega et al. [6]. **Fairness Definition.** Each position in a ranked list carries an inherent position bias that increases with the position number, meaning that even if all items had the same relevance, those at the top of the ranking would receive a lot more attention compared to items at lower ranks. Biega et al. [6] frame this discrepancy as a problem of distributive individual fairness, aiming to achieve equity of attention on the level of individuals, and postulating that the attention an item gets from users should be proportional to their relevance to the query. Assuming that relevance decreases linearly and attention decreases geometrically, there is necessarily a discrepancy between the attention loss of a candidates and their decrease in relevance.

This work proposes a post-processing algorithm that optimizes equity of user attention with a constrained overall relevance loss. (Note that this is in contrast to Singh and Joachims [42], discussed in Section 3.3.4, that constrains fairness and optimizes for relevance.) Because a single ranking cannot be fair according to their definition, the authors propose an approach in which unfairness is mitigated *over time*. For a pair of well-suited

candidates a and b, attention enjoyed by them should be equalized over m rankings $\boldsymbol{\tau}_{1...m}$:

$$\frac{\sum_{i=1}^{m} att(\boldsymbol{\tau}_i, a)}{\sum_{i=1}^{m} U(\boldsymbol{\tau}_i, a)} = \frac{\sum_{i=1}^{m} att(\boldsymbol{\tau}_i, b)}{\sum_{i=1}^{m} U(\boldsymbol{\tau}_i, b)} \tag{15}$$

Hence, unfairness is measured as the accumulated difference between the attention enjoyed by the candidates and their relevance:

$$\text{unfairness}(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_m) = \sum_{a=1}^{n} \left| \sum_{i=1}^{m} att(\boldsymbol{\tau}_i, a) - \sum_{i=1}^{m} U(\boldsymbol{\tau}_i, a) \right| \tag{16}$$

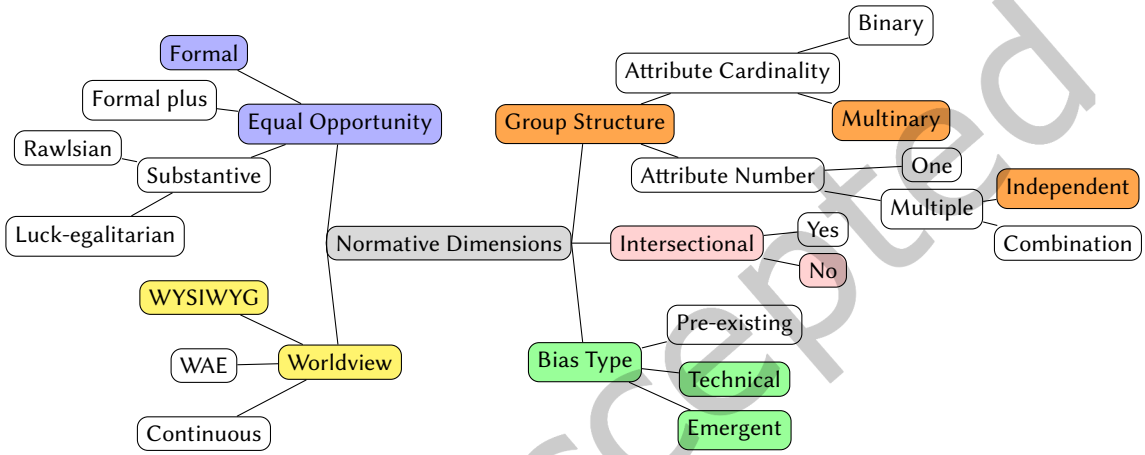

Fig. 15. Summary of the normative values encoded by Biega et al. [6]. This is the only method in this survey that explicitly addresses emergent bias.

Insights. The authors relate their work to the notion of individual fairness by Dwork et al. [15] and treat group fairness, which they call "equality of attention," as a special case, when utility distributions are uniform across all rankings: $U(\boldsymbol{\tau}, a) = U(\boldsymbol{\tau}, b), \forall a, b$. Note that this understanding of group fairness can not account for biases and errors in the data that are manifesting differently across groups, meaning that the error for the protected group might be high, while the data for the privileged group may be error-free. This method is consistent with formal EO. This work's definition of group fairness corresponds therefore explicitly with a WYSIWYG worldview and contrasts with most of the other definitions in the literature. In addition to an explicit focus on technical bias, this method can also address emergent bias, which may result from online learning algorithms that learn through user feedback and click data. Figure 15 and Table 2 summarize our analysis.

Algorithm. The fairness definition is implemented as a constrained optimization problem (which is then translated into an Integer Linear Program) in which unfairness is minimized subject to constraints on the maximum NDCG-loss in an online manner. This means that the algorithms allows unfairness minimization over time, and candidates can enter and leave the system at any point. The algorithm reorders a given ranking such that unfairness is minimized given the cumulative attention and relevance distribution seen so far:

$$\begin{aligned} &\text{minimize} \quad \sum_{a} \left| \sum_{i=1}^{l-1} att(\boldsymbol{\tau}_i, a) + att(\boldsymbol{\tau}_l, a) - \left(\sum_{i=1}^{l-1} U(\boldsymbol{\tau}_i, a) + U(\boldsymbol{\tau}_l, a) \right) \right| \\ &\text{subject to} \quad \text{NDCG@k}(\boldsymbol{\tau}_l, \boldsymbol{\tau}_{l^*}) \geq \theta \end{aligned} \tag{17}$$

where $\boldsymbol{\tau}_l$ is the current, $\boldsymbol{\tau}_{l^*}$ is the reordered ranking, and θ is the quality constraint, with higher values corresponding to more fairness.

As this method measures the attention each item received over time given their relevance, we have to consider several rankings to illustrate its effect. As attention is defined through position bias too, just as in [42], let us come back to Figure 9. It is likely that the first few rankings will look like the ranking $\hat{\boldsymbol{\tau}}_1$ in Figure 9, with candidate f ranked right below candidate e. However, at some point, the attention received by candidate f will be much lower than attention received by its neighbor e, because the relevance of f decreases linearly, while its attention decreases geometrically. If this happens, then candidates e and f will be swapped in the next ranking produced by the model.

Experiments. The experiments are run on two real datasets, AirBnB [**?**] and StackExchange [**?**], and a synthetic dataset, each with two different models for attention gains by position: (1) geometric attention decrease by position, and (2) the first position gets all attention.

- **Synthetic Data:** Experiments are set up with three different relevance score distributions (uniform, linear, exponential) and the aforementioned two attention models. In all cases, the algorithm shows periodic behavior under lab conditions, meaning that every x rounds, it brings unfairness to 0. Furthermore, fairness is not a steady state but will grow with each new ranking for each individual item.
- **AirBnB:** For these experiments rankings were created from AirBnB apartment listings in Hongkong (4529 items), Boston (3944 items) and Geneva (1728 items) under two scenarios — (1) always using the same query, (2) using different queries. The results verify that the difference between the distributions of attention and relevance is substantial in real world datasets, and whether unfairness can be effectively mitigated is highly dependant on the dataset at hand. In all cases, unfairness did not increase over time only when no utility constraint was given ($\theta = 0$.
- ***StackExchange:*** *The experiments show that individual subjects appear in relatively few result rankings, leading to an extended fairness amortization time.*

3.4 Discussion of the Normative Framework Mapping

In this section, we have seen how technical decisions lead to different value frameworks that are supported by a method. In some cases these differences are quite obvious: those method that explicitly incorporate a notion of utility *into their fairness objective*, namely Biega et al. [6], Lahoti et al. [26], Singh and Joachims [43], and the disparate treatment and disparate impact definition of Singh and Joachims [42], generally lean towards the WYSIWYG worldview, with $CS \approx OS$, and are consistent with formal EO. In contrast, methods that explicitly exclude a utility measure from the fairness definition (Geyik et al. [19], Zehlike et al. [57], Zehlike and Castillo [58], Zehlike et al. [59], and the demographic parity definition of Singh and Joachims [42]), generally lean towards the WAE worldview and substantive EO. Additionally, some methods explicitly allow continuous interpolation between these two worldviews, either by introducing a sliding parameter, or by allowing a range of values for the fairness constraints (Geyik et al. [19], Zehlike et al. [57], Zehlike and Castillo [58], Zehlike et al. [59]).

However, the devil is in the details, and even though some methods appear to make very similar technical assumptions, a minor difference in design choices can lead to tremendous differences in their underlying value frameworks. Let us take FA*IR (Sec. 3.3.1) and LinkedIn (Sec. 3.3.2) as examples. The technical choices of these two algorithms appear to be extremely similar: both receive a vector of minimum proportions for the protected groups, and reorder a ranking such that candidates from all groups are shown throughout the ranking according to these minimum proportions. Neither method incorporates utility into the fairness constraints, and both seem to continuously adopt worldviews between WYSIWYG and WAE. However, Geyik et al. [19] explicitly decided to sort all protected candidates based on their utility and always pick the highest-scoring candidate whenever ties in the proportion would allow different choices. In the same situation, Zehlike et al. [57] explicitly pick the next

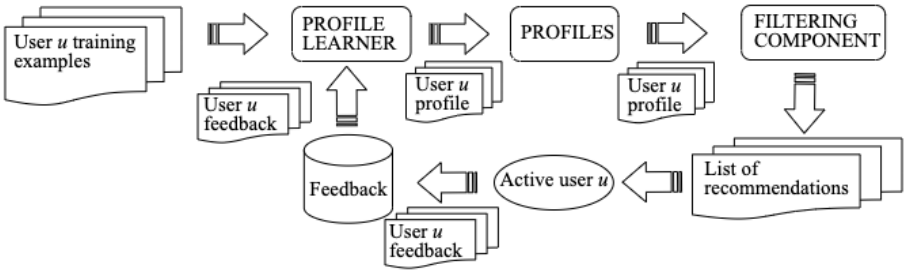

Fig. 16. Principle components of a recommender system. A profile learner determines user profiles from training examples (upper right) which are stored in a profile database. When the user issues a query, the filtering component processes the general search result, and returns a list of user specific recommendations. The interaction of the user are collected as feedback and stored in a feedback database. This data is used in turn to update the model of the profile learner. Image taken from De Gemmis et al. [13].

candidate based on a probability distribution, which again is not based on any utility measure. Geyik et al. [19] incorporate a utility notion for protected candidates through a backdoor, which is why we classified it as leaning towards WYSIWYG and formal EO, even though it is technically very similar to FA*IR. This design choice also has tremendous consequences for individuals who experience intersectional discrimination.

Being able to map mathematical formulas to normative dimensions, and thus to identify normative value frameworks is crucial for several reasons: First, it facilitates the decision on which method to choose under different application scenarios. Obviously one wants to use methods whose value frameworks match the ones of the application at hand. However, we have seen that some methods are not easy to analyse for their values because important details are missing. This was particularly true for the method of Beutel et al. [4], where an explicit formula on how to calculate the click through rate is needed to come up with a mapping. Second, our comprehensive analysis identified research gaps in terms of value frameworks. For example, we could only identify one method, Biega et al. [7], which explicitly addresses the problem of emergent bias. At the same time, the method is consistent with WYSIWYG and with formal EO. Thus, to the best of our knowledge, there is currently no method that explicitly handles emergent bias and is consistent with supports WAE. Finally, in contrast to the work presented in Part I of this survey, no methods discussed in this part of the survey explicitly handle intersectional discrimination. To the best of our knowledge, no such methods have been developed to-date in the learning-to-rank literature.

4 FAIRNESS IN OTHER DOMAINS: RECOMMENDATION AND MATCHING

Today many web applications with search functionality also implement a recommendation system that provides personalized search results to the user with items dedicated precisely to their interest. The main objective of recommender systems is to facilitate transactions between multiple stakeholders in which personalization plays a key role and therefore fairness issues for more than one group of participants have to be considered. RecSys tasks can be grouped into three different types: finding good items that meet a user's interest, optimizing the utility of users, and predicting ratings of items for a user [25]. They often consist of multiple models, must balance multiple goals and are difficult to evaluate due to extreme and skewed sparsity [4]. Examples for such systems are user recommendations in web shops or on streaming platforms. Figure 16 shows the functional principle of a recommender system. A profile learner determines user profiles from training examples (upper right) which are stored in a profile database. When the user issues a query, the filtering component processes the general search

result (and additional items to be displayed), and returns a list of user specific recommendations. The interaction of the user with the result are collected as feedback and stored in a feedback database. This data is used in turn to update the model of the profile learner.

Burke [9] identifies three different stakeholders in recommender systems, namely the consumers, producers and the platform itself. They further identify three different types of recommender systems, distinguished by the respective stakeholders that are considered for fairness. When considering fairness for the recommended items, they speak of producer-fairness, while fairness for the users of the system is considered under the term consumer-fairness. In a system that satisfies consumer fairness the disparate impact of recommendation on protected classes of consumers has to be taken into account, while the fairness of outcomes is not considered for producers. Producer fairness regards the producer side of the system, but not the consumer side, e.g. to ensure market diversity and avoid monopoly domination. At the same time the producers are a passive system component that do not seek out recommendation opportunities, but instead wait for users to request recommendations from the platform. The third fairness constraint simultaneously takes producers and consumers into account and has to be applied when protected groups exist among both stakeholders.

In our paper's language consumers would be called users $\mathcal{U}$ und producers would be called candidates C.

Recommender systems often use rankings to present the most suitable candidates to their users, and the question of the fairness for all stakeholders has been raised from different perspectives. We therefore give a brief overview on the topic of fairness for recommender systems, and refer the reader to several recent tutorials and surveys of fairness in recommendation and matchmaking systems for additional details: Ekstrand et al. [16] work out which algorithmic fairness concepts from classification and scoring do and do not translate to information access scenarios. Li et al. [31] summarize foundations and algorithms for fairness in recommendation systems. Chen et al. [12] outline seven types of biases in recommendation that typically stems from user behavioral data, and provide a taxonomy of existing work on de-biasing. Gao and Shah [17] bridge socio-technical terminologies and metrics to important recommendation concepts such as relevance, novelty, diversity, bias, and fairness.

While maintaining the same level of detail as in other parts of this survey, we will discuss three papers that highlight different directions within fairness in recommender systems. In Section 4.1, we discuss Kamishima et al. [25] to showcase a theoretical contribution. In Section 4.2, we discuss Mehrotra et al. [34] to demonstrate a countermeasure against the familiarity bias feedback loop in the music streaming platform Spotify. Finally, in Section 4.3 we discuss Sühr et al. [48], as an example of work on two-sided fairness for repeated matches.

Before diving in, we also briefly discuss other notable work on fairness in recommendation and matching that do not strongly connect to ranking. An early work on user fairness in recommendations by Leonhardt et al. [28] shows that methods that only focus on diversification on results can lead to discrimination among users. Rastegarpanah et al. [36] present a bias mitigation approach by augmenting the input to an unfair recommendation model with additional "antidote data," as an analogy to work studying data poisoning. Wang et al. [51] study the compositionality of fairness definitions for recommender systems and provide a set of conditions under which fairness of individual models does indeed compose. Sonboli and Burke [44] examine fairness in situations where membership in protected groups is not given a priori, but must be derived from the data itself. Additionally, the authors propose a localized understanding of fairness in cases where global system properties are insufficient to identify protected groups. Deldjoo et al. [14] propose a probabilistic framework based on generalized cross entropy to measure fairness for a given recommendation model, instead of understanding fairness as the divergence of a system from some form equality. Liu et al. [32] propose a fairness method to produce recommendations in the microlending domain that are fair for the borrowers and attentive to individual lender preferences. Ge et al. [18] account for the dynamic nature of recommender systems and consider long-term fairness by proposing a fairness-constrained reinforcement learning algorithm, which models the recommendation problem as a constrained Markov decision process. Sonboli et al. [45] describe the results of an exploratory interview study that investigates user perspectives on fairness-aware recommender systems and on techniques for enhancing their transparency.

Zhu et al. [63] introduce a tensor-based fair recommendation system that preserves the benefits of using matrix factorization for recommendation, while increasing fairness by excluding sensitive features and their latent influence on non-sensitive ones. Li et al. [30] study unfair recommendation performance disparities for active vs. inactive users, since those who actively interact with the platform produce larger amounts of data for collaborative filtering approaches.

In this section we will use the terms users and consumers, as well as candidates and producers interchangeably to demonstrate that these concepts are analogous to each other, and to illustrate the similarities between the previously described methods, and those that follow in this section.

All methods presented in this section can be considered in-processing approaches.

4.1 Recommendation Independence

Fairness Definition. The work of Kamishima et al. [25] introduces a concept of fairness as *recommendation independence*: an unconditional statistical independence between a recommendation outcome and a specified piece of information. In other words, predictions of ratings should not be based on some previously specified feature. This is formalized as a regularization-based approach that can deal with the encoding of sensitive information in the first and second moments of the distributions, which means that independence shall be given in terms of the mean *and* the standard deviation of the distribution of predicted user ratings. A binary sensitive feature will be specified by user or manager.

Let us assume that the random variable $\mathbf{U}$ represents users, $\mathbf{X}$ represents items, $\mathbf{A}$ represents sensitive features, and Y represents ground truth ratings. The i-th instance of the training dataset $\mathcal{T}_{\text{train}}$ is a 4-tuple $(\mathbf{U}_i, \mathbf{X}_i, \mathbf{A}_i, Y_i)$. The rating predictions $\hat{Y}$ are calculated using a modified loss function in which an independence term shall be maximized, meaning that the higher $\text{ind}(\hat{Y}, \mathbf{A})$, the less statistically dependent are $\hat{Y}$ and $\mathbf{A}$:

$$\sum_{(\mathbf{U}_i, \mathbf{X}_i, \mathbf{A}_i, Y_i) \in \mathcal{T}_{\text{train}}} \text{loss}(Y, \hat{Y}(\mathbf{U}_i, \mathbf{X}_i, \mathbf{A}_i)) - \eta \, \text{ind}(\hat{Y}, \mathbf{A}) + \text{reg}(\Theta)$$

The method can therefore be seen as an in-processing approach.

Kamishima et al. [25] give three ways in which independence can be measured, all aiming to produce a rating model in which the distributions of predicted ratings are indistinguishable for different values of the sensitive feature:

(1) **Mean Matching:** The means of two normal distributions shall match

$$-\left(\frac{\mathbb{S}^{(0)}}{N^{(0)}} - \frac{\mathbb{S}^{(1)}}{N^{(1)}}\right)$$

where $\mathbb{S}^{(\mathcal{G})}$ is the sum of predicted ratings per group, and $N^{(\mathcal{G})}$ is the number of training items per group.

(2) **Distribution Matching:** The similarity between two distributions $Pr[\hat{Y}|A = 0]$ and $Pr[\hat{Y}|A = 1]$ is measured in terms of the negative Bhattacharyya-distance:

$$\frac{1}{2} \ln \left(\frac{2\sqrt{\mathbb{V}^{(0)}\mathbb{V}^{(1)}}}{\mathbb{V}^{(0)} + \mathbb{V}^{(1)}} \right) - \frac{\left(\frac{\mathbb{S}^{(0)}}{N^{(0)}} - \frac{\mathbb{S}^{(1)}}{N^{(1)}}\right)^2}{4(\mathbb{V}^{(0)} + \mathbb{V}^{(1)})}$$

where $\mathbb{V}^{(\mathcal{G})}$ is the variance of the training items per group.

(3) **Mutual information:** The degree of statistical independence is quantified by a differential entropy function for normal distributions:

$$-I(\hat{Y}; \mathbf{A}) = -(H(\hat{Y}) - H(\hat{Y}|\mathbf{A}))$$

where $H(\hat{Y}) = \frac{1}{2} \ln 2\pi e \mathbb{V}$ and $H(\hat{Y}|\mathbf{A}) = \frac{1}{2} \ln 2\pi e \mathbb{V}^{(s)}$.

Insights. Although this is not explicitly stated, the goal of equalizing rating distributions is consistent with the WAE worldview. Further, since the fairness definitions do not contain a measure of utility or merit, and under the assumption that the goal is to equalize access to opportunity over a lifetime, the approach is consistent with substantive EO. Because of explicit conditioning on group membership, we map this approach to luck-egalitarian EO. The definition implicitly addresses the problem of pre-existing biases. Depending on the choice of the protected feature (e.g., popularity), it may also be capable to address emergent bias. Figure 17 summarizes this analysis.

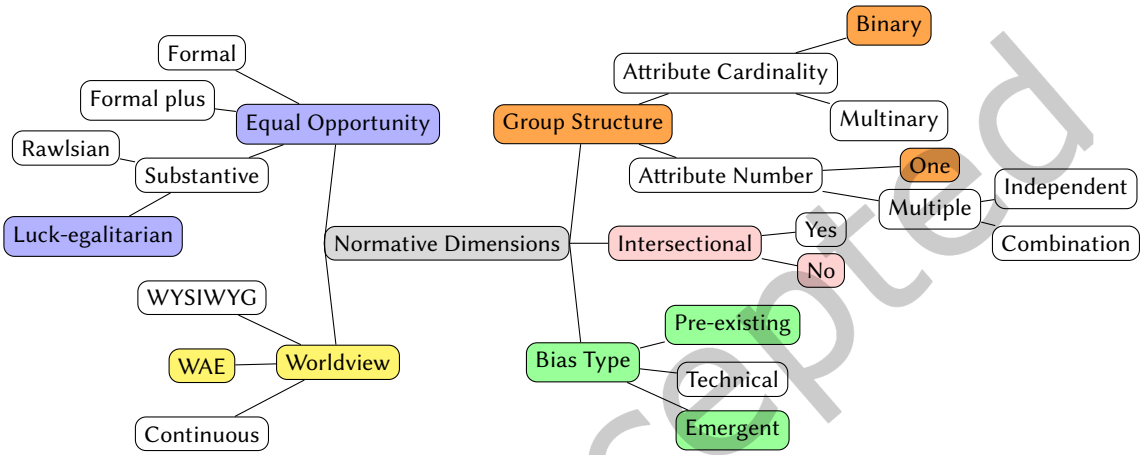

Fig. 17. Summary of the normative values encoded in the fairness definitions by Kamishima et al. [25].

Algorithm. The algorithm is implemented using probabilistic matrix factorization to predict ratings:

$$\hat{Y}(\mathbf{U}, \mathbf{X}, \mathbf{A}) = \mu^{(\mathcal{G})} + b_{\mathbf{U}}^{(\mathcal{G})} + c_{\mathbf{X}}^{(\mathcal{G})} + \mathbf{p}_{\mathbf{U}}^{(\mathcal{G})\top} \mathbf{q}_{\mathbf{X}}^{(\mathcal{G})}$$

where μ, $b_{\mathbf{U}}$ and $c_{\mathbf{X}}$ are global, per-user and per-item bias parameters respectively and $\mathbf{p}_{\mathbf{U}}$ and $\mathbf{q}_{\mathbf{X}}$ are K-dimensional parameter vectors, which represent the cross effects between users and items. Said parameters are learned by minimizing the loss function, where the loss is expressed as a L_2-norm:

$$\sum_{(\mathbf{U}_i, \mathbf{X}_i, \mathbf{A}_i, Y_i) \in \mathcal{T}_{\text{train}}} (Y_i - \hat{Y}(\mathbf{U}_i, \mathbf{X}_i, \mathbf{A}_i))^2 - \eta \operatorname{ind}(Y, \mathbf{A}) + \operatorname{reg}(\Theta)$$

All independence measures are differentiable and can hence be optimized efficiently using conjugate gradient methods.

Experiments. The experiments are done on three datasets with different sensitive features:

- The ML1M **Movielens dataset** [20] contains 1M movie items. Two sensitive features where chosen for two separate settings: (1) whether the movie's release year was before 1990, and (2) the user's gender. The first setting investigates fairness on the side of the producer, whereas the second relates to a fairness concern w.r.t. the users of the system.
- The **Flixster dataset** [21] is also a dataset on movie recommendations and contains almost 10M entries. The popularity of an item was chosen as the protected feature: movies that received the most user ratings (top 1%) belong to one group, while all movies that received fewer ratings belong to the other group.

- The **Sushi dataset** [24] contains around 50,000 data points of users and their preferences when ordering Sushi from 25 different restaurants. Three different choices of sensitive features where investigated: (1) whether the user was a teen, (2) gender: male or female, and (3) whether the type of sushi was seafood.

The performance evaluation uses mean absolute error, while the independence (i.e., fairness) evaluation uses the Kolmogorov-Smirnoff test. The latter evaluates the area between two empirical cumulative distributions and shall be close to zero for high fairness. The experiments compare all three independence measures to each other, and show that independence can be achieved in all cases, with a small cost in accuracy.

4.2 Familiarity Bias and Superstar Economics

Fairness Definition. The work of Mehrotra et al. [34] addresses the problem of disparities in exposure of items to users due to the combination of two factors: the pre-existing familiarity of the user with certain items (someone who likes action movies will likely know Tom Cruise) and the current recommendation strategies of two-sided markets ("superstar economics"). The fact that recommendation systems optimize for relevance can lead to a lock-in of popular and relevant suppliers, and thus cause many suppliers at the tail end of the exposure distribution to struggle to attract consumers. This may, in turn, lead to a dissatisfaction of such suppliers with the marketplace.

Mehrotra et al. [34] aim to understand the interplay between relevance, fairness and satisfaction, and investigate consumer relevance and supplier fairness on a music streaming platform. The paper introduces a notion of multinomial group fairness, which requires that the content shown to users be spread well across the wide long-tailed popularity spectrum, rather than focusing on a small subset of popular artists. From the popularity distribution of all artists, K bins of equal size are created and artists are grouped into these bins depending on their popularity:

$$\Psi(s) = \sum_{i=1}^{K} \sqrt{|t_j|_{\forall t_j \in P_i \cap T(s)}}$$

where P_i is the set of artists that belong to popularity bin i, s is the recommended set, and $T(s)$ is the collection of artists in the set s with t_j being the j-th track in s. This definition rewards sets that are *diverse* in terms of the represented popularity bins and, as per the current definition, *fair* to different popularity bins of suppliers. There is more benefit to selecting an artist from a popularity bin that is not yet represented. As soon as at least one artist is selected from a bin, other artists from the same bin start having diminishing gain owing to the square root function.

The paper then presents three different policies with $\phi(.)$ being a relevance and $\psi(.)$ a fairness measure:

(1) A weighted combination of relevance and fairness:

$$s_u^* = \arg\max_{s \in S_u} (1-\beta)\phi(u,s) + \beta\psi(s)$$

where S_u is the collection of all sets pertinent to the user u.

(2) A probabilistic combination of relevance and fairness, where the weighting factor $\beta \in [0,1]$ decides on whether to recommend a set based on relevance or fairness:

$$s_u^* = \begin{cases} \arg\max_{s \in S_u} \psi(s), & \text{if } p < \beta \\ \arg\max_{s \in S_u} \phi(u,s), & \text{otherwise} \end{cases}$$

where $p \in [0,1]$ is a randomly generated number.

(3) A guaranteed relevance term to ensure that the minimum relevance is β:

$$s_u^* = \arg\max_{s \in S_u} \psi(s) \;\; s.t. \;\; \phi(u,s) \geq \beta$$

They further investigate different affinities of users to fairness; for example, that some users only want to listen to one particular artist, while other users are more flexible. This affinity is measured as the difference between the satisfactions of a user when recommended relevant content vs. fair content. The paper states that a fairness recommendation policy should be adaptive to this affinity ξ_u, and therefore redefine the second fairness policy as:

$$s_u^* = \begin{cases} \arg\max_{s \in S_u} \psi(s), & \text{if } \xi_u \geq 0 \\ \arg\max_{s \in S_u} \phi(u, s), & \text{otherwise} \end{cases}$$

Another version redefines the first policy to use the z-scored affinity, denoted as $\hat{\xi}_u$:

$$s_u^* = \arg\max_{s \in S_u} (1 - \hat{\xi}_u)\phi(u, s) + \hat{\xi}_u \psi(s)$$

Because all definitions are modifying the objective function of the learning algorithm, the method can be classified as in-processing approach.

Insights. The method is addressing a combination of pre-existing and technical bias, where the popularity bias can be seen as pre-existing, and its reinforcement by the recommender system as technical bias. Furthermore, because of the temporal nature of recommender systems, proposed methods may also address emergent bias w.r.t. popularity shifts in the future.

The work does not explicitly express the authors' beliefs about the mapping g from construct space CS to observable space OS. That being said, the fairness definition aims to equalize the exposure of different artist independent of their popularity and can therefore be associated with the WAE worldview. Further, because of a natural connection to emergent bias, the assumption of equalizing opportunity over the lifetime is also very natural. For these reasons, the methods can be associated to substantive EO, and specifically with luck-egalitarian EO, which would acknowledge different distributions of popularity in construct space. However, it is not clear how an artist's popularity relates to their true underlying effort at all, as it is mainly driven by the *users* of the system, rather than by the artists themselves. Figure 18 summarizes our analysis.

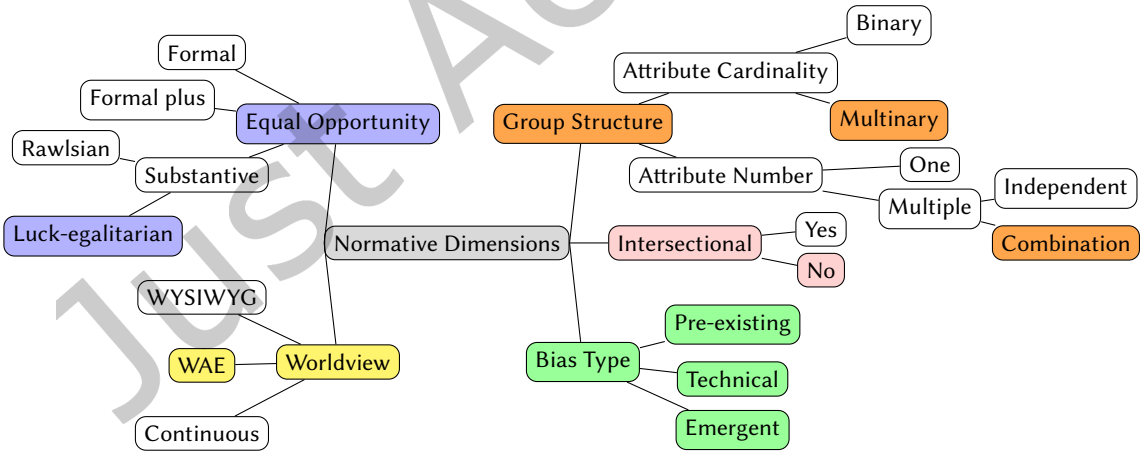

Fig. 18. Summary of the normative values encoded by Mehrotra et al. [34].

Algorithm. The algorithm is implemented as a combinatorial contextual bandit problem with the following consecutive interactions between the customers and the recommendation system:

(1) The system observes context $\mathbf{X}$ drawn from a distribution of contexts $\mathbf{P}(\mathbf{X})$.

(2) Based on $\mathbf{X}$, the system chooses the sets s to recommend to the user.
(3) A reward function $Y \in [0, 1]$ is drawn from the distribution $\mathbf{P}(Y|\mathbf{X}, s)$ that expresses user satisfaction.
(4) The system maximizes Y under different fairness policies.

As estimating user satisfaction is not easy, because it relies on user feedback which is hard to estimate offline, the recommender system is modeled as a stochastic policy π that specifies a conditional distribution $\pi(a|\mathbf{X})$. The value $V(\pi)$ of a policy is the expected reward (i.e., the user's satisfaction) if an action a is chosen under that policy. This value is to be estimated for a new policy π^* given logged training data, using the inverse propensity score (IPS) estimator, which is provably unbiased:

$$V_{\text{offline}} = \sum_{\forall(\mathbf{X},a,Y_a,p_a)} \frac{Y_a \mathbb{1}(\pi^*(\mathbf{X}) = a)}{p_a}$$

with p_a being the propensity scores for an action a that was randomly chosen from the space of all possible actions under context $\mathbf{X}$, and $\mathbb{1}(\pi^*(x) = a)$ being the set indicator function that evaluates to 1 if the action selected by the target policy matches the logged training data. With this, a metric of user satisfaction can be computed as:

$$Y_{\pi^*(x)} = \mathbb{E}_a \left[\frac{Y_a \mathbb{1}(\pi^*(\mathbf{X}) = a)}{p_a} \right]$$

Experiments. The experimental part contains a trade-off analysis between user relevance and fairness of sets, which shows that only very few highly relevant sets also achieve high scores on fairness. This means that a recommendation system that solely optimizes for user relevance will not automatically lead to fair and diverse sets for the suppliers, and, in turn, that two-sided marketplaces have to optimize for both users and suppliers to satisfy both.

Evaluation is conducted on a proprietary dataset from Spotify with 400K users, 49K artists, and 5K sets (i.e., playlists). When maximizing relevance $\phi(u, s)$ only, the results show the highest user satisfaction, whereas when optimizing for fairness only, user satisfaction drops by 35%. When optimizing for both, user satisfaction increases with an increase of the weight that is given to the relevance objective. However satisfaction does not drop significantly up to a fairness weight of 20%, such that policies could easily increase fairness up to that point without trading user satisfaction significantly. The guaranteed relevance policy yields the best satisfaction results in absolute numbers, and shows a linear trend of user satisfaction with increasing weights for relevance. However, for the same levels of β, this policy achieves the least average fairness scores compared to the other two policies. This means that the usage of the interpolating fairness policy and the probabilistic fairness policy is preferable in situations where less fairness shall be traded for higher relevance values. When evaluating the adaptive policies, the results highlight that personalizing the recommendation policy and adapting based on user level affinity is better than globally balancing relevance and fairness. Interestingly, adaptive policies lead to a relatively high fairness mean compared to global policies, while at the same time increasing the overall user satisfaction. This is another example where the general assumption that a trade-off between fairness and relevance is a necessary evil, is shown not to hold.

4.3 Fairness in Two-sided Markets

Fairness Definition. The work of Sühr et al. [48] gives a case study of a two-sided matching platform, namely, a ride-hailing platform such as Uber. This paper discusses fairness in a platform performing repeated matches of providers (drivers) and consumers (riders) over time. Fairness is seen in terms of fair distribution of driver's income, given their active time in the system. It is explicitly considered over time, because a single matching does not have a significant long-term effect on the life of the people that are matched, and because income equity can be amortized over a longer period, such as a week or a month. Note that the paper speaks explicitly about riders

and drivers, but in a broader sense those can be seen as users $\mathcal{U}$ of the system and candidates C to be matched to them.

Customer utility $U_{\mathcal{U}}$ is described as a customer $\mathcal{U}_b$'s waiting time, which is approximated using the negative distance d of a driver C_a to them:

$$U_{\mathcal{U}}(b, a) = -d(\mathcal{U}_b, C_a)$$

Driver utility is described as the income a driver receives from transporting a customer, which is approximated using the distance from the customer's pick-up location to their destination, reduced by the distance the driver has to travel in order to arrive at the respective pick-up location:

$$U_C(a, b) = d(\mathcal{U}_b, \text{dest}(\mathcal{U}_b^t)) - d(\mathcal{U}_b, C_a)$$

The paper then defines two fairness concepts. First, the authors introduce **parity fairness**: over time, the sum of received utility shall be (almost) equal for all drivers in C:

$$\sum_a \sum_b |U_{C_a}^T - U_{C_b}^T| < \epsilon$$

with $U_{C_a}^T$ being the total utility that driver a received until matching round T. Second, the authors define **proportional fairness**: over time, the sum of received utility normalized by their active driving time shall be equal for all drivers

$$\sum_a \sum_b \left| \frac{U_{C_a}^T}{\Lambda_{C_a}^T} - \frac{U_{C_b}^T}{\Lambda_{C_b}^T} \right| < \epsilon$$

where $\Lambda_D^T(j)$ is the total amount of time a driver has been active on the platform until T.

Insights. The work does not talk about any protected groups and instead tries to equalize the (hourly) wage of all drivers in the system. Also, the drivers' effort is not considered other than in the sense of their active time, and, as such, it does not make sense to associate this method with a particular equality of opportunity framework. The assumption behind this work is that driving skills are essentially the same and therefore all drivers should be paid equally. It is an extreme case of the WAE worldview in which not only do all groups have the same qualification distribution, but the absolute qualification values are the same for each individual. The biases addressed are of a technical and emergent nature. Emergent bias would result in the sense that the drivers' locations are shaped by the platform, and driver may be concentrated in certain hot spots, while some locations remain less frequented, potentially disadvantaging riders who which to be picked up near those locations. Figure 19 summarizes our analysis.

Algorithm. The method defines a two-sided optimization objective to minimize the difference of the utilities of drivers (resp. customers) as compared to the maximum utility gained by any driver (resp. customer) up until the previous matching round:

$$\sum_a \sum_b \lambda \cdot \left| \max_{a'} U_{C_{a'}}^{T-1} - \left(U_{C_a}^{T-1} + M_{a,b}^T \cdot U_C^T(a, b) \right) \right| + (1 - \lambda) \cdot \left| \max_{b'} U_{\mathcal{U}_{b'}}^{T-1} - \left(U_{\mathcal{U}_b}^{T-1} + M_{a,b}^T \cdot U_{\mathcal{U}}^T(b, a) \right) \right|$$

where $M_{a,b}^T$ is 1 if driver C_a is matched to customer $\mathcal{U}_b$ in round T and 0 otherwise, and λ is a hyper-parameter to continuously interpolate between producer and consumer fairness. This objective is translated into an integer linear program.

Experiments. Experiments are performed on a dataset from a ride hailing platform in an Asian city consisting of 1462 registered drivers, and measure income inequality using the generalized entropy index (GEI).. As passengers

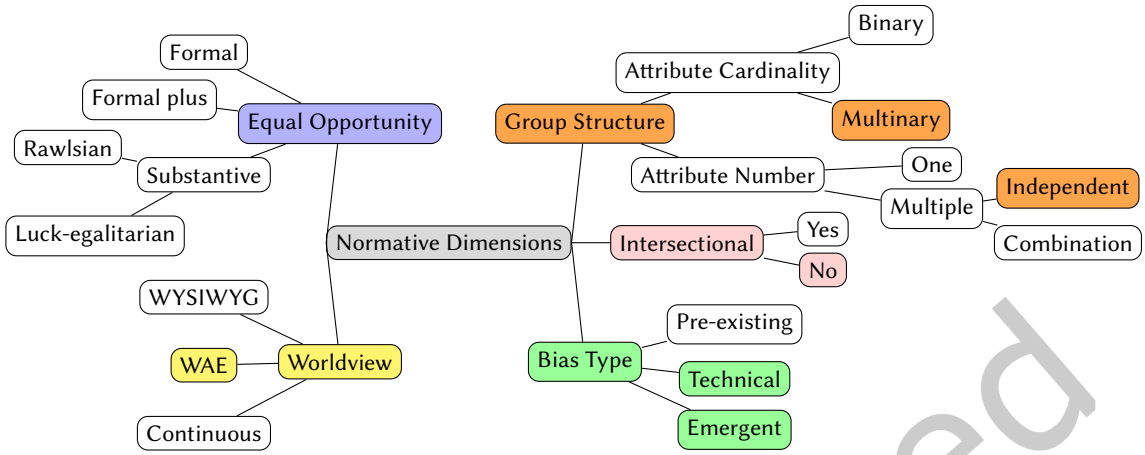

Fig. 19. Summary of the normative values encoded by Sühr et al. [48].

are not registered, every request is handled using a unique job id, with 231,268 jobs in total. The analysis of the dataset shows that supply exceeds demand by an order of magnitude, and driver income is, hence, a scarce resource. Then different matching strategies are compared to each other w.r.t. their effects on driver and customer utility:

- **Nearest driver first** is a simple objective for low customer waiting times. It yields the highest *average* driver utility, but has the largest discrepancies in terms of driver income (i.e., a high GEI).
- **Worst-off driver first** yields almost equal driver income, but lowest customer utility. It also shows undesired effects, such as lowering hourly wages for drivers that have long active periods, because whenever a new driver joins the system, they have the lowest possible income so far, namely, zero.
- **Two-sided optimization** achieves better results both in terms of income equality for drivers and in terms of customer waiting time, because drivers happen to be placed in better positions for subsequent requests. However, it is not clear whether this result is an artifact of the dataset, or if it points to a general property of the strategy. In any case, this finding confirms that, just as is the case for rankings [58] and for two-sided platforms [34], fairness does not always come at the cost of utility. In fact, optimizing for fairness can sometimes increase system utility.

5 FRAMEWORKS AND BENCHMARKS

In this section, we survey several frameworks and benchmarks that focus on fairness in ranking, and are relevant to both score-based ranking methods from Part I (Section 5) and to supervised learning methods from Part II (Section 3).

5.1 Fair Search

FairSearch [61] is an open source API to provide fair search results, which is designed as stand-alone libraries (supported in Python and Java) and plugins of Elasticsearch (supported in Java). Users can run FairSearch together with their own datasets once these have been formatted as required. FairSearch is designed for learning-to-rank and information retrieval tasks, where the input is queries and documents with relevance scores to each query. For an unseen query, FairSearch outputs a ranked list of relevant documents for it.

FairSearch implements two algorithms to guarantee fair ranking as search results: an in-processing technique named as DELTR (see details in Section 3.2.1) and a post-processing one called FA*IR (details in Section 3.3.1) For the support of DELTR, FairSearch provides an off-line wrapper to train a fair ranking model using DELTR that is later uploaded and stored in Elasticsearch as plugins. For FA*IR, FairSearch applies FA*IR algorithm to rerank the search results provided by Elasticsearch and presents it as final search results to users. The implementation of FairSearch can be found in https://github.com/fair-search.

5.2 TREC Fair Ranking Track

The Text REtrieval Conference (TREC)[4], an annual conference operated by the U.S. National Institute of Standards and Technology (NIST) to support information retrieval research, has a track for fair ranking that evaluates systems according to how well they fairly rank documents, starting from 2019. Each year's Fair Ranking track defines 1-2 information retrieval tasks that ask participants to build systems to address. The tasks are provided with a corpus of documents as training data and a set of queries for evaluation. The participants submit their system's responses to these evaluation queries, and the task organizers score each submission based on some pre-defined measurements that are also provided to the participants with the tasks. Thus, the Fair Ranking track could be seen as a benchmark that allows the comparison of different methods to guarantee fairness in information retrieval tasks.

The Fair Ranking track 2019 [5] focuses on re-ranking academic abstracts given a query. The objective is to fairly represent relevant authors from several, undisclosed group definitions. Note that these groups can be defined in a variety of ways, and the definition of groups is not provided by the task organizers and will be a part of evaluation to test the robustness of the results for various group definitions. The Fair Ranking track 2020 also focuses on scholarly search and fairly ranking academic abstracts and papers from authors belonging to different groups. The 2021 focused on fairly prioritising Wikimedia[5] articles for editing to provide a fair exposure to articles from different groups.

5.3 Ranking Facts

Ranking Facts [56] is a Web-based application that generates a "nutritional label" for rankings. The nutritional label is made up of a collection of visual widgets that implement research results on fairness, stability, and transparency for rankings, and that communicate details of the ranking methodology, or of the output, to the end user. Ranking Facts is designed for score-based rankings, where rankers are specified by users as input.

Ranking Facts supports tabular datasets as inputs. Users are asked to upload a dataset as in CSV format and specify a score-based ranking function. For fairness and diversity concerns, users are also required to specify an attribute in the uploaded data as the sensitive one to define group membership. Then, Ranking Facts generates a nutritional label of the generated ranking. The fairness widget in a produced label implements three definitions of group fairness: whether a fair ranking shows proportional representation, encodes pairwise comparison, or satisfies the definition by FA*IR (see details in Section 3.3.1). Ranking Facts can be accessed at https://dataresponsibly.github.io/tools/. The code is available at https://github.com/DataResponsibly/RankingFacts.

6 DISCUSSION AND RECOMMENDATIONS

In Part I and Part II of the survey, we discussed the principal functioning of fair ranking methods, and made explicit the technical choices they make and the value frameworks they encode. In this section we discuss our insights and draw a set of recommendations for the evaluation of fair ranking methods.

[4]https://trec.nist.gov
[5]https://www.wikimedia.org

Our recommendations are aimed at data science researchers and practitioners. With these recommendations we aim to establish best practices for the development, evaluation, and deployment of fair ranking algorithms, and to avoid potentially harmful uninformed transfer of methods from one application domain to another.

Recommendation 1: Make Values and Context of Use Explicit. Different application scenarios require different value frameworks. The classification frameworks we presented in this paper are meaningful if the application scenario is concerned with aspects of *distributive justice*. However, even if a situation requires distributive justice to be taken into account, the goods or benefits to be distributed play a key role in determining which framework should be applied.

For example, college admissions (educational opportunity) may require a different interpretation of fairness than hiring or user rating prediction in online shops (economic opportunity). To avoid algorithmic solutionism, users of fair rankers must first clearly articulate their own moral goals for a ranking task, and choose a fairness-enhancing method that is consistent with their goals. To facilitate this choice, the fairness concepts behind a fair rankers must be made explicit by their creators.

Values are rarely made explicit in fair ranking papers. A reader looking to adopt a method to their application context will often use the experiments section of a paper to decide whether the method will "work" for them. We often see experimental sections in which all available datasets, corresponding to vastly different contexts of use—from recidivism risk prediction, to credit scoring, to college admissions, to matching platforms like AirBnB—are used to show performance of a method, but without an explanation as to why the dataset was selected, other than that it was available and items in it have scores on which to rank. For example, the COMPAS dataset [2] is often used in experiments for papers that propose methods for distributive justice [55, 57], yet the dataset was collected for a legal decision making task and so this use is out of scope. We caution against this practice and recommend that, when designing their experiments, the authors should carefully substantiate the appropriateness of using a proposed method on a particular dataset in the context of a specific ranking task. This substantiation should be made by mapping the method and the task to a value framework.

Recommendation 2: Surface Normative Consequences of Technical Choices. Algorithmic rankers are complex technical objects, and many implicit and explicit choices go into their design. In this paper we discussed an important technical dimension of ranker design, namely, the representation of group structure: how many sensitive attributes a ranker handles, and whether these are binary or multinary. This technical choice in turn impacts what type of discrimination a fair ranker can help address (e.g., on one or on several sensitive attributes), and whether it can address intersectional concerns and, if so, what specific concerns are in scope (e.g., representation constraints on intersectional groups, differences in score distributions, or imbalanced loss in utility). We deliberately discussed intersectional discrimination under the heading of mitigation objectives, rather than presenting it as a purely technical choice, and we recommend that designers of fair rankers explicitly discuss both their technical choices, and what consequences they have for applicability.

Another important technical dimension is where in the processing pipeline bias mitigation is applied (recall Figure 6 on page 11 in the first part of this survey). Pre-processing methods have the advantage of early intervention on pre-existing bias. The advantage of in-processing methods is that they do not allow a biased model to be trained. The advantage of post-processing methods is that they provide a guaranteed share of visibility for protected groups. However, post-processing methods are be subject to legal debates in the US because of due process concerns that may make it illegal to intervene at the decision stage (e.g., Ricci v. DeStefano [49]). In the EU, post-processing methods can be used if other methods fail to comply with EU anti-discrimination law. We recommend that the designers of fair rankers substantiate the appropriateness of their technical choice based on the context of use, on for which their method was designed, as well as on the region of use to avoid legal pitfalls.

Recommendation 3: Draw Meaningful Comparisons. Additional research is needed to understand how methods with opposing worldviews, conflicting understandings of equality of opportunity, and different addressed biases can be compared in a meaningful way. For example, in their experiments the authors of Fair-PG-Rank [43] compare its results to those of DELTR [58], but as their measure of disparate exposure is vastly different and the bias they are focusing on is not the same, it is not clear what conclusion to draw from such comparisons. Making the values and the context of use explicit will go a long way towards helping design meaningful experimental comparisons between methods, rather than mechanically comparing apples to oranges.

7 CONCLUSION

In this survey we gave an extensive overview of the state-of-the-art literature of fair ranking in the domains of score-based and supervised learning-based ranking. We introduced important dimensions along which to classify fair ranking methods, mapping their assumptions and design choices to the normative values they incorporate. We outlined the technical details of all methods, presented commonalities and differences, and categorized each technical choice by its implicit values within the normative dimensions. We discussed implications of normative choices and gave recommendations for researchers on how to make such choices in their work explicit.

Most fair ranking methods are concerned with the concept of *distributive justice*, as they aim to fairly distribute the visibility in a ranking among the candidates. Our focus on distributive justice allowed for the mapping between the worldviews and the equality of opportunity concepts in the framework we proposed. However, this mapping is only meaningful in a distributive context and most likely cannot be transferred to a different setting. In the future we hope to see work that relates to other concepts of justice, such as *procedural justice*, which is concerned with the fairness and transparency of a decision making *process*, and is therefore particularly important in legal decision making. It will be interesting to study whether fairness-enhancing methods designed for concerns of distributive justice can be transferred to the context of procedural justice in a meaningful way.

Another interesting direction to classify fair ranking methods within distributive justice contexts is to understand the properties of ranking scores with respect to different indexes of advantage. Commonly, the score models a candidate's *potential utility* to the user of the ranking, which stems from welfarism/utilitarianism and, as such, incorporates an idea of satisfaction and preference [40]. In contrast, Rawls judges the goodness of a distribution in terms of so-called *primary goods*. It is important to explore the implications of these different conceptions of advantage, and, crucially, understand whether they can be combined in a common fairness objective.

8 ACKNOWLEDGEMENTS

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. This research was supported in part by NSF Awards No. 1934464, 1916505, and 1922658.

REFERENCES

[62]]AirBnBData AirBnB. [n. d.]. AirBnB. https://insideairbnb.com

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2016).

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

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking Through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. ACM, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[5] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2019. Overview of the TREC 2019 Fair Ranking Track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings.*

[6] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 405–414.

[7] 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.

[8] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

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

[10] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. ACM, 129–136.

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

[12] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240* (2020).

[13] Marco De Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2015. Semantics-aware content-based recommender systems. In *Recommender systems handbook*. Springer, 119–159.

[14] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogin, and Tommaso Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction* (2021), 1–55.

[15] 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.

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

[17] Ruoyuan Gao and Chirag Shah. 2020. Counteracting Bias and Increasing Fairness in Search and Recommender Systems. In *Fourteenth ACM Conference on Recommender Systems*. 745–747.

[18] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards Long-term Fairness in Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

[19] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.

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

[21] Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*. 135–142.

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

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

[24] Toshihiro Kamishima. 2003. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 583–588.

[25] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation independence. In *Conference on Fairness, Accountability and Transparency*. 187–201.

[26] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1334–1345.

[27] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439* (2019).

[28] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*. 101–102.

[29] Hang Li. 2014. *Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00607ED2V01Y201410HLT026

[30] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021*. 624–632.

[31] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. SIGIR.

[32] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 467–471.

[33] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Evaluation in information retrieval. *Introduction to information retrieval* 1 (2008), 188–210.
[34] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.
[62]]COMPASData ProPublica. [n. d.]. Correctional Offender Management Profiling for Alternative Sanctions. https://github.com/propublica/compas-analysis
[36] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 231–239.
[62]]GermanCreditData UCI Machine Learning Repository. [n. d.]. German Credit. https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/
[38] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*. 521–530.
[62]]SATData SAT. [n. d.]. SAT. https://www.qsleap.com/sat/resources/sat-2014-percentiles
[40] Amartya Sen. 1980. Equality of what? *The Tanner lecture on human values* 1 (1980), 197–220.
[62]]LSACData Law School Admission Council Research Report Series. [n. d.]. LSAC national longitudinal bar passage study. https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/LawStudents
[42] 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.
[43] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. *arXiv preprint arXiv:1902.04056* (2019).
[44] Nasim Sonboli and Robin Burke. 2019. Localized fairness in recommender systems. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 295–300.
[45] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.
[62]]StackData StackExchange. [n. d.]. StackExchange. https://stackexchange.com/
[62]]EngineeringData Engineering students. [n. d.]. Engineering students. https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/EngineeringStudents
[48] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 3082–3092.
[49] Supreme Court of the United States. 2009. Ricci v. DeStefano (Nos. 07-1428 and 08-328), 530 F. 3d 87, reversed and remanded. https://www.law.cornell.edu/supct/html/07-1428.ZO.html.
[62]]W3CData TREC. [n. d.]. W3C Experts. https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/TREC
[51] Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H Chi, Jilin Chen, and Alex Beutel. 2021. Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 436–444.
[52] Linda F Wightman and Henry Ramsey. 1998. *LSAC national longitudinal bar passage study*. Law School Admission Council.
[62]]XINGData Xing. [n. d.]. XING. https://github.com/MilkaLichtblau/xing_dataset
[62]]YahooData Yahoo. [n. d.]. The Yahoo Webscope Program. https://webscope.sandbox.yahoo.com/
[55] 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.
[56] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, 1773–1776.
[57] 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.
[58] Meike Zehlike and Carlos Castillo. 2018. Reducing disparate exposure in ranking: A learning to rank approach. *arXiv preprint arXiv:1805.08716* (2018).
[59] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2017. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery* (2017), 1–38.
[60] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with multiple protected groups. *Information Processing & Management* 59, 1 (2022), 102707.
[61] Meike Zehlike, Tom Sühr, Carlos Castillo, and Ivan Kitanovski. 2020. FairSearch: A Tool For Fairness in Ranked Search Results. In *Companion Proceedings of the Web Conference 2020*. 172–175.

[62]]YowData Yi Zhang. [n. d.]. Yow News Recommendation. https://www.younow.com
[63] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1153–1162.