Reducing Disparate Exposure in Ranking: A Learning To Rank Approach

Meike Zehlike
Humboldt Universität zu Berlin
Max-Planck-Institut for Software Systems
meikezehlike@mpi-sws.org

Carlos Castillo
Universitat Pompeu Fabra
chato@acm.org

ABSTRACT

Ranked search results have become the main mechanism by which we find content, products, places, and people online. Thus their ordering contributes not only to the satisfaction of the searcher, but also to career and business opportunities, educational placement, and even social success of those being ranked. Researchers have become increasingly concerned with systematic biases in data-driven ranking models, and various *post-processing* methods have been proposed to mitigate discrimination and inequality of opportunity. This approach, however, has the disadvantage that it still allows an unfair ranking model to be trained.

In this paper we explore a new *in-processing* approach: DELTR, a learning-to-rank framework that addresses potential issues of *discrimination* and *unequal opportunity* in rankings at training time. We measure these problems in terms of discrepancies in the *average group exposure* and design a ranker that optimizes search results in terms of relevance *and* in terms of reducing such discrepancies. We perform an extensive experimental study showing that being "colorblind" can be among the best or the worst choices from the perspective of relevance and exposure, depending on how much and which kind of bias is present in the training set. We show that our in-processing method performs better in terms of relevance and exposure than a pre-processing and a post-processing method across all tested scenarios.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; **Top-k retrieval in databases**; • **Applied computing** → *Law, social and behavioral sciences*;

KEYWORDS

Ranking, Algorithmic Fairness, Disparate Impact

ACM Reference Format:
Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366424.3383534

1 INTRODUCTION

Ranked search results have become the main mechanism by which we find content, products, places, and people online. These rankings are typically constructed to provide maximum utility to searchers, by ordering items by decreasing probability of being relevant [18]. However, when the items to be ranked represent people, businesses, or places, ranking algorithms have consequences that go beyond immediate utility for searchers. Researchers have become increasingly concerned with various systematic biases [10] against socially-salient groups, caused by historic and current discriminatory patterns making their way into data-driven models. A common element in this line of research is the presence of a historically and currently disadvantaged *protected group*, and the concern of *disparate impact*, i.e., loss of opportunity for the protected group independently of whether they are (intentionally) treated differently. In the case of rankings, a natural way of understanding disparate impact is by considering differences in exposure [21] or inequality of attention [3], which translate into systematic differences in access to economic or social opportunities.

Disparate exposure in rankings. A number of issues, sometimes appearing jointly, call for reducing disparate exposure in information retrieval systems. First, there can be a situation in which minimal differences in relevance translate into large differences in exposure across groups [3, 21], because of the large skew in the distribution of exposure brought by positional bias [14]. Second, there can be a legal requirement that requires protected elements to be given sufficient visibility among the top positions in a ranking [7, 25]. Third, there can be systematic discrepancies in the way in which documents are constructed, as in the case of certain sections in online resumes, which are completed differently by men and women [1]; these discrepancies may in turn systematically affect ranking algorithms. Fourth, there can be systematic differences in the way ground truth rankings have been generated due to historical discrimination and/or annotator bias. These issues point to two conceptually different goals: *reducing inequality of opportunity* (as defined by O'Neill [15]) and *reducing discrimination* (as defined by Roemer [19], chapter 12). Equality of opportunity seeks to correct a historical or present disadvantage for a group in society. Non-discrimination seeks to allocate resources in a way that does not consider irrelevant attributes.

Fairness-aware methods. These methods can be classified into *pre-*, *in-* and *post-processing* approaches, where pre-processing methods seek to mitigate discriminatory bias in training data, in-processing methods learn a bias-free model, and post-processing methods re-rank output items [12]. For rankings, several post-processing methods have been presented in the literature [3, 7, 21, 25]. Yet the post-processing approach has several limitations. First, the idea inherently suggests that there is *always* a trade-off between an optimally *fair* and an optimally *relevant* ranking, because a presumably "exact" model produces a "relevant" ranking that is then reordered to meet fairness constraints. Yet our experiments

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383534

reveal that reducing bias against a protected group can increase relevance (Section 6.2). Second, a post-processing procedure still allows an unfair ranking model to be trained on biased features and later deployed. To achieve a fair outcome the only possibility using post-processing is to apply a predefined anti-discrimination policy that hard-codes fairness constraints and potentially ignores relevance judgments. In-processing methods can instead learn to ignore the protected features as well as their proxies. Pre-processing methods do not allow a biased model, yet our experiments show that creating an unbiased training set is not trivial and may easily lead to reverse discrimination. In summary, we make the following contributions:

(1) **Listwise Fairness:** We propose a new metric for fairness in rankings that operates on the concept of disparate exposure. We use this to define the first listwise learning-to-rank (LTR) approach, named DELTR, that is concerned with *reducing disparate impact* at training time.

(2) **New Datasets:** We perform extensive experiments on two different ranking tasks: expert search in a document retrieval setting, and ranking students by predicted performance. Our experiments comprise three real-world datasets, of which two are newly introduced (Section 5).

(3) **Non-Discrimination vs. Equal Opportunity:** Our experimental descriptions draw a clear distinction between scenarios in which we seek to reduce discrimination, and situations in which we want to enhance equal opportunity, which is yet missing in the algorithmic fairness literature.

(4) **Study on Colorblindness:** As stated by Dwork et al. [9], being "colorblind" on discriminatory training data, i.e. merely ignoring protected attributes, can be a bad idea, because non-protected attributes serve as proxies for the protected ones [4]. In our experiments we analyze in which cases colorblindness yields the best results, and in which it is among the worst results both in terms of relevance and fairness. We also explain how these cases are related to contribution 3, and show that DELTR performs well in terms of fairness and relevance in all tested scenarios.

(5) **FA*IR as Pre-Processing Approach:** We demonstrate a pre-processing approach for fairness in rankings by applying a post-processing method, FA*IR [25], to our training data before the learning routine starts. These experiments show two interesting insights: (i) it is not easy to produce fair training data, because discrimination may be embedded in all attributes, and a truly bias-free dataset is hard to obtain; and (ii) re-ordering items in a "fair" way can lead to significant performance decline and even to reverse discrimination.

2 RELATED WORK

Fairness in ranking is concerned with a sufficient presence, a consistent treatment, and a proper representation of different groups across all ranking positions [6]. At a high level, this line of research has the goal of producing rankings based on relevant characteristics of items, in which items belonging to the protected group are not under-represented or systematically relegated to lower ranking positions [24]. Singh and Joachims [20] introduce the concept of *exposure* of a group, based on empirical observations that show that the probability that a user examines an item ranked at a certain position, decreases rapidly with the position. We will use this concept to present a new evaluation metric that measures exposure as the average probability of a group to be ranked in the top position. Previous works on fair rankings [3, 7, 20, 23, 25] have been concerned with creating a fairness-aware ranking from a given set of scores, and can be considered *post-processing* approaches—they are given a ranking and re-rank elements to achieve a desired objective. In contrast, our approach DELTR is *learning-based* as it extends ListNet [5], a well-known listwise LTR framework. It constitutes the first listwise *in-processing* approach to reduce discrimination and inequality of opportunity in rankings, because it learns a ranking function with an additional objective that reduces disparate exposure. While the recently proposed pairwise approach by Beutel et al. [2] cares about *disparate treatment*, our listwise method directly optimizes the actual exposure a protected group would get, and is hence concerned with *disparate impact*. Also we do not take user feedback into account, as it constitutes an additional source of unconscious biases, that we want to study separately.

3 BACKGROUND: LISTNET IN A NUTSHELL

We consider a set of queries Q with $|Q|= m$ and a set of documents D with $|D|= n$. Each query q is associated with a list of candidate documents $d^{(q)} \subseteq D$, where each document is represented as a feature vector $x_i^{(q)}$. For each query the list of feature vectors $x^{(q)}$ is associated with a list of judgments: $x^{(q)} \rightarrow y^{(q)}$. The standard objective then is to learn a ranking function f that outputs a list $\hat{y}^{(q)}$ of new judgments $\hat{y}_i^{(q)}$ for each feature vector $x_i^{(q)}$. Ideally, the function f should be such that the sum of the differences (or losses) L between the training judgments $y^{(q)}$ and the predicted judgments $\hat{y}^{(q)}$ is minimized: $\min\left(\sum_{q\in Q} L\left(y^{(q)}, \hat{y}^{(q)}\right)\right)$.

As rankings are combinatorial objects, the naive approach to find an optimal solution for L leads to exponential execution time in the number of documents. Hence, instead of considering an actual permutation of documents, Cao et al. [5] only focus on the probability for a document $d_i^{(q)}$ to be ranked in the top position:

$$P_{\hat{y}^{(q)}}\left(d_i^{(q)}\right) = \frac{\phi\left(\hat{y}_i^{(q)}\right)}{\sum_{j=1}^{n}\phi\left(\hat{y}_j^{(q)}\right)} \tag{1}$$

with $\phi : \mathbb{R}_0^+ \longrightarrow \mathbb{R}^+$ being an increasing strictly positive function. The top-one-probabilities form a probability distribution of judgments over $d^{(q)}$. By setting $P_{y^{(q)}}(x_i^{(q)})$ to be the top-one-probabilities of the ground truth and $P_{\hat{y}^{(q)}}(x_i^{(q)})$ to be those of the predictions, Cao et al. [5] measure the loss between $y^{(q)}$ and $\hat{y}^{(q)}$ using the Cross Entropy metric:

$$L\left(y^{(q)}, \hat{y}^{(q)}\right) = -\sum_{i=1}^{|d^{(q)}|} P_{y^{(q)}}(x_i^{(q)}) \log\left(P_{\hat{y}^{(q)}}(x_i^{(q)})\right) \tag{2}$$

4 DELTR: DISPARATE EXPOSURE IN LEARNING TO RANK

For our listwise fairness approach we assume that the retrieved items belong to two distinct social groups (such as men and women, or majority and minority ethnicity), and that one of these groups is *protected* [17]. At training time, we are given an annotated set consisting of queries and ordered lists of items for each query. At testing time, we provide a query and a document collection, and expect as output a list of top-k items from the collection that

should be relevant to the query, and additionally should not exhibit disparate exposure.

Disparate Exposure. We assume that items in D belong to two different groups, which we denote by G_0 for the non-protected group, and G_1 for the protected group. Items in the protected group have a certain protected attribute, such as belonging to an underprivileged group. As argued in Section 1, the protected group may, due to various causes including historic discrimination or erratic data collection procedures, have a significant disadvantage in the training dataset. This is likely to cause a model to predict rankings with a large discrepancy in exposure, and not only to reproduce but reinforce discrimination and unequal opportunities for already disadvantaged groups.

To define a measure of "unfairness" we borrow the definition of Singh and Joachims [21] on exposure of a document d in a ranked list generated by a probabilistic ranking P, and adapt it for top-one-probabilities (eq. 1) to match ListNet's accuracy metric:

$$\text{Exposure}\left(x_i^{(q)}|P_{\hat{y}^{(q)}}\right) = P_{\hat{y}^{(q)}}\left(x_i^{(q)}\right) \cdot v_1 \tag{3}$$

where v_1 is the *position bias* of position 1, indicating its relative importance for users of a ranking system [13]. Hence, the average exposure of documents in group G_p with $p \in \{0, 1\}$ is

$$\text{Exposure}(G_p|P_{\hat{y}^{(q)}}) = \frac{1}{|G_p|} \sum_{x_i^{(q)} \in G_p} \text{Exposure}(x_i^{(q)}|P_{\hat{y}^{(q)}}) \tag{4}$$

Finally, we adapt the first definition of equal exposure in [21], *demographic parity*, which ensures that the average exposure across items from all groups is equal. With this we can now introduce an unfairness criterion measured in terms of disparate exposure:

$$U(\hat{y}^{(q)}) = \max\left(0, \text{Exposure}(G_0|P_{\hat{y}^{(q)}}) - \text{Exposure}(G_1|P_{\hat{y}^{(q)}})\right)^2 \tag{5}$$

Note that in contrast to [21], using the squared hinge loss gives us a metric that prefers rankings in which the exposure of the protected group is not less than the exposure of the non-protected group, *but not vice versa*. This means that our definition will optimize only for relevance in cases where the protected group already receives as much exposure as the non-protected group.

We note that other fairness objectives can be used as long as they can be optimized efficiently (e.g., are differentiable), and that the definition in Equation 5 can be easily extended to multiple protected groups by considering average or maximum difference of exposure between a protected group and the non-protected one.

Formal Problem Statement. Having formalized an accuracy measure L (eq. 2) and a listwise fairness measure U, we can now combine these two into a fair loss function L_{DELTR}. Specifically, we seek to minimize a weighted summation of the two elements, controlled by a parameter $\gamma \in \mathbb{R}_0^+$:

$$L_{DELTR}\left(y^{(q)}, \hat{y}^{(q)}\right) = L\left(y^{(q)}, \hat{y}^{(q)}\right) + \gamma U\left(\hat{y}^{(q)}\right) \tag{6}$$

with larger γ expressing preference for solutions that focus on reduction of disparate exposure for the protected group, and smaller γ expressing preference for solutions that put emphasis on the differences between the training data and the output of the ranking algorithm. The parameter γ depends on desired trade-offs between ranking utility and disparate exposure that are application-dependent. To set it, we looked at the ratio between L and U and used this as $\gamma_{\texttt{small}}$. For γ_{large} we increased γ_{small} by an order of magnitude. We remark that, even if γ is set very high DELTR only increases fairness until exposure for both groups is equal. We confirmed this with synthetic experiments in two different settings: one where all non-protected items appeared at the top positions, and one where all protected items were followed by all non-protected ones. In the first setting, increasing values of γ lead to more exposure of the protected group and items are put to higher positions. However DELTR does not over-compensate and moves protected items only as long as exposure is not equal across groups. In the second case DELTR behaves like a standard LTR algorithm.

Optimization. For the ranking function to infer the document judgments we use a linear function $f_\omega(x_i^{(q)}) = \langle \omega \cdot x_i^{(q)} \rangle$ [5], and Gradient Descent to find an optimal solution for L_{DELTR}. We can now rewrite the top-one-probability for a document (eq. 1) and set ϕ to an exponential function, which is strictly positive, increasing and convenient to derive:

$$P_{\hat{y}^{(q)}(f_\omega)}(x_i^{(q)}) = \frac{\exp(f_\omega(x_i^{(q)}))}{\sum_{k=1}^{n} \exp(f_\omega(x_k^{(q)}))} \tag{7}$$

To use Gradient Descent we need the derivative of $L_{DELTR}(y^{(q)}, \hat{y}^{(q)})$ which in turn consists of the derivatives of the disparate exposure and accuracy metric respectively.

$$\frac{\partial L_{DELTR}\left(y^{(q)}, \hat{y}^{(q)}\right)}{\partial \omega} = \frac{\partial L(y^{(q)}, \hat{y}^{(q)})}{\partial \omega} + \gamma \cdot \frac{\partial U(\hat{y}^{(q)})}{\partial \omega} \tag{8}$$

5 EXPERIMENTS

In our experiments, we consider three real-world datasets summarized in Table 1. We study *non-discrimination*, through experiments that seek to reduce *biases unrelated to utility* (Sec. 6.1), or biases that originate from *different score distributions at the same relevance level* across social groups (Sec. 6.2). Due to the nature of these biases we do not expect to see a trade-off between search utility and list-wise fairness, as both can be achieved at the same time. In the first case (Sec. 6.1), *excluding* the protected attribute for training will lead to the best result in terms of utility and list-wise fairness. In the second case (Sec. 6.2) we want to explicitly *include* the protected feature to achieve higher utility and less disparate exposure. DELTR can handle both cases without prior knowledge about the underlying bias. Additionally we study *substantive equality of opportunity*, through experiments that seek to reduce biases due to utility differences that pre-exist (Sec. 6.3). We apply DELTR to each dataset with two different values of γ: γ_{large} in which γ is comparable to the value of the standard loss L, and γ_{small} in which it is an order of magnitude smaller. Then we compare the results against several baselines: (i) a "colorblind" LTR approach, which excludes protected attributes during training; (ii) a standard LTR method, which considers them during training; (iii) a post-processing approach that applies LTR and then re-ranks the output; and (iv) a pre-processing approach that modifies the training data.

W3C experts (TREC Enterprise) Dataset. This dataset originates from the expert search task at the TREC 2005 Enterprise Track [8], where an algorithm has to retrieve a sorted list of experts

	W3C Experts (gender)	Engineering Students (high school type)	Engineering Students (gender)	Law Students (gender)	Law Students (race)
Prediction Task	Expertise	Academic performance	Academic perf.	Academic perf.	Academic perf.
Ranking score	Expertise level	Weighted first year average	WFYA	FYA	FYA
#items/query	200	480.6 (ave.)	480.6 (ave.)	21791	19567
#folds	6	5	5	1	1
Queries	Technical topics	Acad. year	Acad. year	Acad. year	Acad. year
#Q_{train}/fold	50	4	4	80%	80%
#Q_{test}/fold	10	1	1	20%	20%
Protected attr.	female	public high school	female	female	black
#protected/query	21.5 (ave.)	167.6 (ave.)	97.6 (ave.)	9537	1282

Table 1: Datasets summary. The law student dataset has only one query, training and test are obtained by an 80/20 split.

for a given topic, given a corpus of e-mails written by possible candidates. While all experts are considered equally expert, we injected a discriminatory pattern in this dataset by sorting the ground truth for each training query in the following order: 1. all male experts, 2. all female experts, 3. all male non-experts, and 4. all female non-experts. This simulates a scenario where expertise has been judged correctly, but training lists have been ordered with a bias against women, placing them systematically below men at the same level of expertise. We computed a series of text-retrieval features for each query-document pair, such as word count and tf-idf scores by usage of the Elasticsearch Learning to Rank Plug-in [16].

Engineering Students Dataset. The dataset contains anonymized historical information from first-year students at a large school in a Chilean university. As qualification features we are given the results of the Chilean university admission test named PSU in categories math, language, and science, their high-school grades, and the number of credits taken in their first year.

Law Students Dataset. This dataset originates from a study by Wightman [22] that examined whether the LSAT (Law Students Admission Test in the US) is biased against ethnic minorities. It contains anonymized historical information from first-year students at different law schools. We use a uniform sample of 10% of this dataset, while maintaining the distribution of gender and ethnicity.

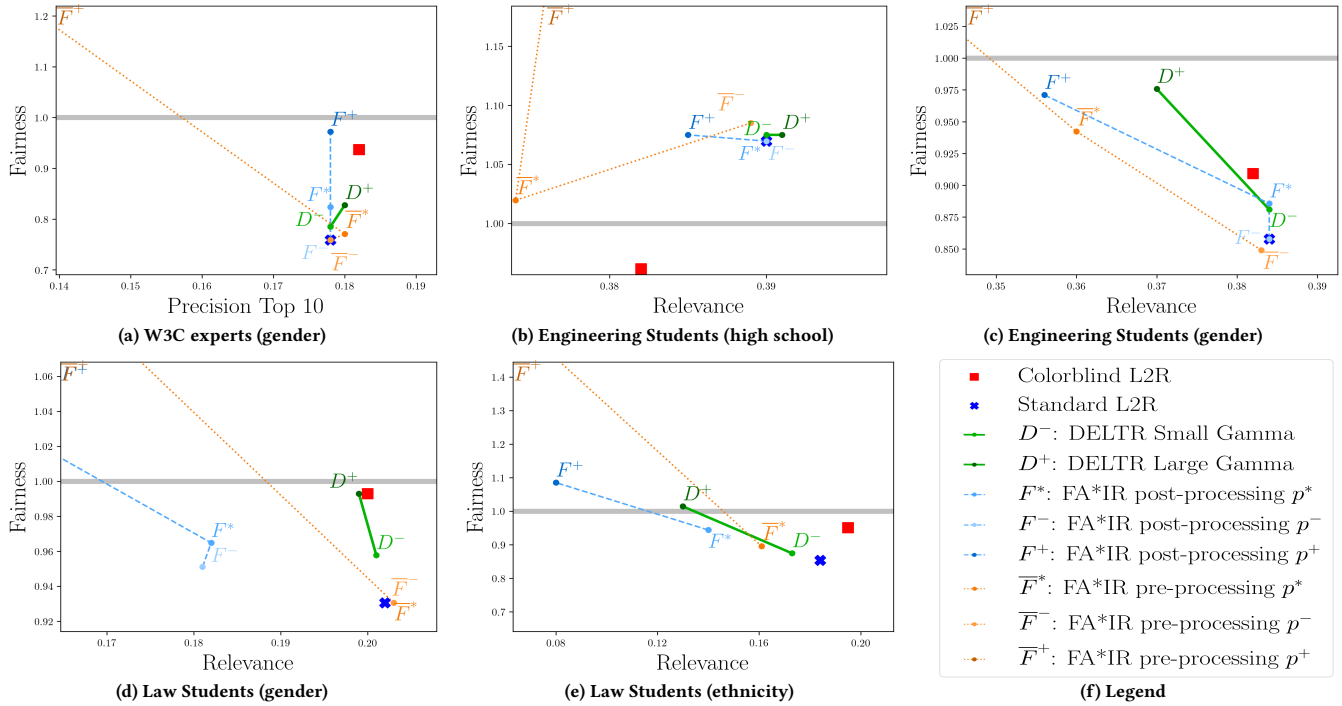
Baselines. We compare DELTR with a small and a large value for γ to *pre-, in-* and *post-processing* approaches. Our *in-processing baselines* constitute (i) ListNet, a standard LTR algorithm [5], which is applied "colorblindly", i.e. over all non-sensitive attributes; and (ii) the same LTR approach in which all attributes are used (including the protected one). In the *pre-* and *post-processing* baselines we apply the algorithm FA*IR [25] to the training data and the predicted rankings of a standard LTR method, respectively. FA*IR is a top-k ranking algorithm that ensures a minimum target proportion p of a protected group at every prefix of a ranking based on a statistical significance test. In our *pre-processing* baseline experiments we process a given training dataset with FA*IR to free the data from potential bias and create fair training data. We use three different values of p, p^* = the ratio of protected candidates in the dataset, $p^+ = p^* + 0.1$ and $p^- = p^* - 0.1$, to show how crucial the right choice of p is, especially in a pre-processing setting. Afterwards we use ListNet [5] to train a ranker over all features, both sensitive and non-sensitive. The *post-processing* baseline also uses ListNet and trains a ranker over all available features, including the protected one. Then FA*IR is applied to the predicted rankings, potentially resulting in a reordering of the items. We use the same parameters p^*, p^+ and p^- as in the pre-processing experiments.

6 EXPERIMENTAL RESULTS

In this section we present the results of each experimental setting, which are depicted in Figure 1 and summarized in Table 2.

6.1 Bias Unrelated to Utility – W3C Experts

Experimental results are shown on Figure 1a, averaged over all folds, using $\gamma_{\texttt{small}} = 20K$, $\gamma_{\texttt{large}} = 200K$, and $p^* = 0.105$, which is the proportion of women in the dataset. In this experiment we expect the "colorblind" approach to achieve the best results, because we injected a strong bias against women *that was completely unrelated to their expertise.* The setting corresponds to a *non-discrimination* case, where we want to exclude the protected feature from training for relevance reasons and we expect to see no trade-off between accuracy and list-wise fairness when optimizing for both. Figure 1a confirms our expectations. Note that we measure utility in terms of precision at ten instead of Kendall's tau, because we want to know which algorithm finds most of the true experts and ranks them accordingly. Colorblind LTR performs best in terms of relevance and achieves almost equal exposure for men and women, by distributing women evenly across rankings. Standard LTR (including the biased protected feature) performs worse in terms of relevance and exposure than most of the other approaches. Indeed, the model discriminates against women based solely on their gender, by placing all women at the bottom of the ranking (not shown), even those that were considered experts in the ground truth. The in-processing approach DELTR reduces the gap in exposure between men and women, and scores best in terms of relevance compared to all other fair algorithms. Post-processing (blue "F") with p^+ achieves better exposure, but leads to a slight over-representation of women at the top-positions, which causes the lower relevance w.r.t. DELTR. When using pre-processing FA*IR (orange "$\overline{F}$") with the intuitive p^*, the model is not de-biased, meaning that this value for p is too low for this setting. However, pre-processing using p^+, which is only slightly larger than p^*, not only increases exposure to the profound detriment of the non-protected group, but also performs significantly worse than all other approaches in terms of relevance (Figure 1a, all dots lying above the gray line in the figures mean that the protected group now receives higher exposure than the non-protected one). These effects of a too small or too large p for all cases of FA*IR, post- *and* pre-processing can be seen in all following



(a) W3C experts (gender)
(b) Engineering Students (high school)
(c) Engineering Students (gender)
(d) Law Students (gender)
(e) Law Students (ethnicity)
(f) Legend

Figure 1: (Best seen in color.) Comparison of relevance and fairness (i.e. exposure) achieved by each approach. The horizontal line indicates equal exposure for both groups. We see that a trade-off between list-wise fairness and relevance is not universal. Instead its presence or absence depends on the concrete underlying bias in the training data (plot 1b vs 1c). In case we observe a trade-off between performance and exposure (plot 1c, 1d and 1e), DELTR mostly outperforms the pre- and post-processing approaches. The plots focus on high-relevance results and settings that obtain substantially lower relevance are omitted. Their approximate position can be inferred from the lines that join settings of the the same approach.

results: a too small p shows no effect on the exposure of the protected group in the rankings. However, a too large p can result in an over-representation of protected elements at the top positions. This may result into inverting the bias, such that non-protected items are now ranked low solely because of their group membership. In contrast on the one hand DELTR always results in better exposure, even if γ is set low. On the other hand it excludes the risk of reverse discrimination by design. This advantage comes from the fact that in-processing methods consider both objectives *simultaneously*. They constantly trade relevance against fairness measures until the best balance is found, while pre- or post-processing approaches examine relevance and fairness measures consecutively and hence the sweet spot must be found manually.

6.2 Bias due to Different Score Distributions – Engineering Students (high school type)

In this experiment, we consider students coming from public high schools as the protected group and those from private high schools as the non-protected. Results appear in Figure 1b ($\gamma_{\texttt{small}} = 100K$, $\gamma_{\texttt{large}} = 5M$ and $p^* = 0.348$, which is the proportion of students from public high schools). The ground truth shows that students from public schools perform worse on average in the admission test, but tend to have higher grades in university than students from private high schools with the same scores. One explanation for this phenomenon is that public schools tend to provide an education of inferior quality compared to private schools in Chile. For achieving the same test scores, students from public schools need to have better academic aptitudes (similar to observations in [11]). This scenario corresponds to achieving *non-discrimination* with different underlying score distributions, while the same ground truth utility exists across social groups. Under these circumstances, *including* the protected attribute will lead to better performance in terms of relevance *and* exposure. We therefore expect the colorblind LTR to be among the worst approaches, and standard LTR to be among the best. The results in Figure 1b confirm our expectations. We see that the colorblind method performs significantly worse than most approaches both in terms of exposure and in terms of relevance. DELTR, given that students from the protected group already receive higher exposure, does not further increase their ranks, preserving the quality of the ranking result (due to the asymmetry of the method). The same is true for FA*IR in pre- and post-processing, in this case with *small* values of p. Recall however that a small p did *not* do the trick in exp 6.1, because those bias' properties were of a different kind. DELTR can handle both types of biases without knowing their nature a-priori. In the post-processing setting FA*IR with p^+ achieves equal exposure ratios as DELTR, but less relevance. In the pre-processing experiment, a too large p-value (p^* and p^+), leads the LTR algorithm to place too much weight on the protected feature, resulting in a strong decline of relevance.

6.3 Achieving Substantive Equal Opportunity

As the remaining three experiments all relate to the same goal of achieving *substantive equal opportunity* [15], we will describe our findings jointly in this section.

Experiment	W3C Experts (gender)		Engineering Students (high school type)		Engineering Students (gender)		Law Students (gender)		Law Students (race)	
	P@10	Fairness	Kendall's Tau	Fairness	K. Tau	Fairness	K. Tau	Fairness	K. Tau	Fairness
Colorblind LTR	0.182	0.936	0.382	0.962	0.382	0.909	0.200	0.993	0.195	0.951
Standard LTR	0.178	0.759	0.390	1.070	0.384	0.858	0.202	0.931	0.184	0.853
DELTR $\gamma_{\texttt{small}}$	0.178	0.785	0.390	1.075	0.384	0.860	0.201	0.958	0.173	0.874
DELTR $\gamma_{\texttt{large}}$	0.180	0.827	0.391	1.075	0.370	0.976	0.199	0.993	0.130	1.014
FA*IR post p^*	0.178	0.824	0.390	1.070	0.384	0.886	0.182	0.965	0.140	0.944
FA*IR post p^+	0.178	0.972	0.385	1.075	0.356	0.971	0.143	1.074	0.080	1.085
FA*IR post p^-	0.178	0.759	0.390	1.070	0.384	0.858	0.181	0.951	–	–
FA*IR pre p^*	0.180	0.770	0.374	1.020	0.360	0.942	0.203	0.931	0.161	0.895
FA*IR pre p^+	0.052	2.058	0.376	1.203	0.307	1.223	0.149	1.186	0.041	1.726
FA*IR pre p^-	0.178	0.759	0.389	1.085	0.383	0.849	0.203	0.931	–	–

Table 2: Experimental results. Relevance is expressed as Kendall's Tau except for the W3C dataset, where we use P@10. In this experiment we want to see all experts in the top positions rather than produce the correct ordering of the entire list. Fairness is measured as the exposure ratio between the protected and the non-protected group. Hence values < 1.0 mean more visibility for the non-protected group, while values > 1.0 mean more visibility for the protected group.

Engineering students (gender). Figure 1c summarizes the results obtained with parameters $\gamma_{\texttt{small}} = 3K$, $\gamma_{\texttt{large}} = 50K$, and $p^* = 0.202$, which is the proportion of women in this dataset.

Law students (gender). Figure 1d summarizes the results with $\gamma_{\texttt{small}} = 3K$, $\gamma_{\texttt{large}} = 50K$ and $p^* = 0.437$, which is the proportion of women in this dataset.

Law students (race). Results appear in Figure 1e using parameters $\gamma_{\texttt{small}} = 1M$, $\gamma_{\texttt{large}} = 50M$ and $p^* = 0.064$, which is the proportion of African-American students. We did not use p^- because it would have been a negative number.

Interpretation. From the ground truth we know for all three experiments that the protected group scores worse than the non-protected one in their admission tests and also worse in terms of academic success after the first year. We therefore expect a trade-off between utility and exposure, if we optimize for more exposure than the protected group should receive based on their "true" performance. This is desirable if one wants to achieve substantive equal opportunity, corresponding to the usage of a disparate impact approach. If we assumed the training data was free of bias and/or mistakes and truly reflects a student's achievements, the colorblind baseline corresponds to a group's true performance. However we expect neither colorblind nor standard LTR to yield equality of exposure, because the protected group's achievements fall behind the non-protected ones in the ground truth. Using the standard LTR baseline, i.e. including the protected feature into the training phase, leads to even better results in terms of accuracy in figures 1c and 1d than colorblind LTR, but causes a significant drop in exposure for the protected group. Interestingly, in Figure 1e, we observe the reverse: including the protected feature leads to a drop both in accuracy and exposure w.r.t. colorblind. We suspect this happens because the distributions of true performances for each group are far apart in the ground truth, which causes the ranker to overshoot the target by putting far too much weight on the protected feature. Note that this constitutes a very different effect than what was described in Section 6.1, and is not further studied here. As before, being an in-processing approach DELTR consistently outperforms the pre- and post-processing baselines, both in terms of accuracy and in terms of list-wise fairness. This means that using DELTR, we lose less relevance for the same exposure achievement in a search result, than when using pre- or post-processing approaches like FA*IR. A too small p again does not show any effects for the mitigation of disparate impact (pre- and post-processing FA*IR with p^- in figures 1c and 1d; and pre-processing FA*IR with p^* in figure 1d). However a too large p can quickly result not only in over-representation of the protected group, but also yields a significant decline in terms of result relevance, with no upper bound. We interpret this as "too many protected candidates that performed poorly being pushed to higher positions", as FA*IR only takes relevance within groups into consideration. In-processing methods like DELTR can not produce over-representative models because they optimize for exposure and accuracy *at the same time*.

7 CONCLUSIONS

LTR models can reproduce and exaggerate discrepancies of the average group visibility from training data. In this paper we presented the in-processing approach DELTR. It extends ListNet with a list-wise fairness objective that reduces the extent to which protected elements receive less exposure. Our experiments showed that this additional objective does not necessarily come with a trade-off in accuracy. On the contrary, aiming for list-wise fairness will *increase* relevance in cases corresponding to *non-discrimination*. We showed that non-discrimination can be achieved by explicitly *excluding* or *including* the protected feature and studied the nature of underlying biases for each case. As it is hard to understand a-priori which bias is present, DELTR provides a convenient approach to handle both situations.

Future work. The parameter γ provides great flexibility but more work is required to provide a systematic way of setting this parameter. We formalized the extension of our list-wise fairness notion to multiple protected groups, but still need to experimentally validate.

Reproducibility. All datasets and code for reproduction are available at https://github.com/MilkaLichtblau/DELTR-Experiments. DELTR is also available as a stand-alone library in Java and Python, as well as a plugin for Elasticsearch at https://github.com/fair-search.

Acknowledgments. Castillo thanks La Caixa project LCF/PR/PR16/11110009 for partial support. Zehlike thanks the MPI-SWS for their support. The authors thank Gina-Theresa Diehn and Pere Urbon for their assistance during this research.

REFERENCES

[1] Kristen M Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton. 2017. Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles Among Recent MBA Graduates. In *ICWSM*. 460–463.

[2] 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, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. 2212–2220. https://doi.org/10.1145/3292500.3330745

[3] 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.

[4] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[5] 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.

[6] Carlos Castillo. 2018. Fairness and Transparency in Ranking. *SIGIR Forum* 52 (12 2018). Issue 2.

[7] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv:1704.06840 (pre-print)* (2017).

[8] Nick Craswell, Arjen P de Vries, and Ian Soboroff. 2005. Overview of the TREC 2005 Enterprise Track.. In *Trec*, Vol. 5. 199–205.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proc. of Innovations in Theoretical Computer Science (ITCS)*. ACM, 214–226.

[10] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347.

[11] Jennifer Glynn. 2019. *Persistence: The Success of Students Who Transfer from Community Colleges to Selective Four-Year Institutions.* Technical Report.

[12] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

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

[14] 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, 4–11.

[15] Onora O'Neill. 1977. How Do We Know When Opportunities are Equal? In *Feminism and Philosophy*. In M. Vetterling-Braggin and F.A. Elliston and J. English (Eds.), Totowa: Rowman and Littlefield, 177–189.

[16] OpenSource Connections. 2017. We're Bringing Learning to Rank to Elasticsearch. https://opensourceconnections.com/blog/2017/02/14/elasticsearch-learning-to-rank/.

[17] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 560–568.

[18] Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* 33, 4 (1977), 294–304.

[19] John E Roemer. 1998. *Equality of opportunity*. Harvard University Press.

[20] Ashudeep Singh and Thorsten Joachims. 2017. Equality of Opportunity in Rankings. In *Workshop on Prioritizing Online Content (WPOC) at NIPS*.

[21] 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.

[22] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).

[23] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proc. of International Conference on Scientific and Statistical Database Management (SSDB)*. ACM, 22.

[24] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A Nutritional Label for Rankings. *arXiv:1804.07890 (pre-print)* (2018).

[25] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *Proc. of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1569–1578.