Counterfactual Representations for Intersectional Fair Ranking in Recruitment

Clara Rus, Maarten de Rijke and Andrew Yates

University of Amsterdam

Abstract
Fairness interventions require access to sensitive attributes of candidates applying for a job, which might not be available due to limitations imposed by data protection laws. In this work we propose using a pre-processing technique to create counterfactual representations of the candidates that lead to a more diverse ranking with respect to intersectional groups. To be compliant with data protection laws we propose to train a model on the fairer representations and apply the model at inference time without having access to the sensitive attributes of the candidates. In experiments on the BIOS dataset, we find this approach can improve the diversity of recommendations at top-ranked positions without harming performance.

Keywords
Fairness, Recruitment, Intersectionality, Ranking

1. Introduction

Recruiters increasingly rely on automatic hiring systems to process the large amount of applications received for a job. We define an algorithmic hiring system to be a candidate recommendation system that recommends a ranked list of candidates to a recruiter given an occupation. Using an objective automatic hiring system one would think that the hiring process is fair. However, such systems encode stereotypes and biases that already exist in the recruitment industry [5, 9, 16, 22], leading to actions that discriminate against minority groups [2, 11, 21].

Considering that the attention of recruiters decreases with the position of the candidate in the rank [13], the candidates at top positions are more likely to be considered for an interview. To combat existing disparities in the recruitment industry and avoid that they are perpetuated by the system, we aim to apply a fairness intervention to obtain a diverse ranking of candidates, in terms of sensitive attributes at top positions. According to the European General Data Protection Regulation (GDPR), access to special categories of sensitive attributes is limited [1, 23]. Exceptions of special sensitive attributes are gender and age. This limits the choice of fairness interventions that can be used in practice in recruitment.

Existing fairness interventions can be categorized as pre-processing, in-processing and post-processing methods [29]. Pre-processing methods aim to debias the data used to represent the candidates and then either re-rank the candidates based on the new representations or use the data to train a model. In-processing methods aim to optimize the recommendation system for both fairness and utility [4, 27]. Post-processing methods are applied on the recommended ranking. They re-rank the candidates based on some minimum and maximum constraints regarding the desired proportion of each sensitive group in the top positions [21, 24, 28]. All methods require that the sensitive attributes of the candidates are known, implying that they are not compliant with GDPR [1], making it hard to use them in practice.

We argue that pre-processing methods can be used to create fairer representations of candidates that can later be used for offline training a recommendation system, which will generate a fair ranking without having access to sensitive attributes of candidates during inference. There are several methods that can be used to debias data. Lahoti et al. [15] create fair representations of candidates that are individually fair, independent of the sensitive attributes; thus, candidates with similar features should be treated the same, regardless of the sensitive attributes. However, the candidate's data could be a proxy to sensitive attributes, thus, the new representations would keep the same distance between groups formed by the proxy data. Adversarial debiasing methods [30] create fair representations from which the adversary should not be able to predict the sensitive attributes. The representations created eliminate as much as possible any information encoded in the features about the sensitive attributes. Yang et al. [25] create counterfactual representations of the data by estimating the causal effects of the sensitive attributes on the features and scores of the candidates, assuming there is a pre-existing bias in the data.

We investigate how the method proposed by Yang et al. [25] performs in a recruitment scenario. We consider this

RecSys in HR'23: The 3rd Workshop on Recommender Systems for Human Resources, in conjunction with the 17th ACM Conference on Recommender Systems, September 18–22, 2023, Singapore, Singapore.

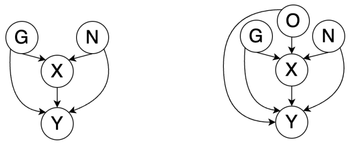
c.a.rus@uva.nl (C. Rus); m.derijke@uva.nl (M. de Rijke); a.c.yates@uva.nl (A. Yates)
0000-0002-0944-5044 (C. Rus); 0000-0002-1086-0202 (M. de Rijke); 0000-0002-5970-880X (A. Yates)

method for the following reasons: (i) it showed promising results even without knowing the sensitive attributes at inference time, (ii) it provides a working framework for intersectional groups, and (iii) it complies with transparency requirements towards recruiters, candidates and audit companies. Unlike adversarial methods, it is easy to explain how the new representations are created and, thus, how the ranking of the candidates is generated. By creating counterfactual representations of candidates, we aim to have more diversity among the top candidates of the ranked list, thus increasing the likelihood of the protected groups to be considered for an interview. Our work focuses on the intersectional groups created by gender and nationality. It is important to take into account intersectional groups, as candidates belonging to multiple protected groups are more likely to be discriminated [24]. We also check whether training a model offline on the counterfactual representations can lead to a diverse ranking without having access to sensitive attributes at inference time. Our **main finding** is that we need to explicitly model the occupation as the bias direction varies across occupations. We also show that by training a model on counterfactual representations the diversity of the ranking is improved and the performance of the model is not affected.

2. Counterfactual Representations

We consider the task of ranking candidates who applied to an occupation listing given a score that represents how well they fit it. Assuming there is a pre-existing bias in the features and scores of the candidates, we aim to create fairer representations by applying the method proposed in [25].

The method uses as input a causal model describing the data and the effects of the sensitive attributes on the data. A causal graph is a directed acyclic graph (DAG) where nodes represent variables, and directed edges between nodes represent causal relationships. A directed edge from node A to node B indicates that variable A causally influences variable B. Figure 1 shows two possible causal models that can be used to represent the data. *Model 1* contains the following nodes: sensitive attributes, G (gender) and N (nationality), non-sensitive attributes of the candidates (X), and the utility score used to rank the candidates for a given occupation (Y), with edges from the features to the scores, and from the sensitive attributes to the features and the scores of the candidates. *Model 2* contains the same set of nodes and edges, and additionally a node (O) representing the occupation with edges to the features and the scores. By adding the occupation node the model captures variations of the direction of bias across occupations.



(a) Model 1 (b) Model 2

Figure 1: Causal models describing the data with sensitive attributes gender (G) and nationality (N), non-sensitive attributes (X), utility scores (Y), and occupation (O).

To estimate the causal effects of sensitive attributes on the data, we have to determine a reference group towards which we want to transform the candidates in a counterfactual world. The idea is to replace the values of sensitive attributes with reference values and propagate the changes in the graph to compute values of the counterfactual features and scores. The method estimates the total causal effect of intersectional sensitive attributes on the score. It estimates the direct effect and the indirect effect mediated by the non-sensitive attributes, which are called mediators. The causal effects are estimated using the mma R package [26], which performs mediation analysis with multiple mediators. To estimate the causal effects we propose three scenarios: do not model the occupation and apply *Model 1* on the whole data (NoOccupation), apply *Model 1* on the data corresponding to each occupation (SingleOccupation), and apply *Model 2* on the whole data by also specifying a reference occupation (ModelOccupation).

After computing the causal effects of the sensitive attributes on the data, which represent the bias encoded in the data, one can compute the counterfactual representations. These are computed by changing the observed representations according to the causal estimates of the sensitive attributes. Counterfactual representations can be used to create a new ranking based on the counterfactual scores, or they can be used to train a model. Such a model can be used at inference time to predict the rank position of a candidate given the counterfactual representations of the candidate or the original representations. For the first option one needs access to sensitive attributes, which, according to the GDPR [1], is not possible in practice. We propose to use the second option: to train a model on counterfactual representations and apply it to the original representations without access to sensitive attributes at inference time.

3. Experimental Setup

The BIOS dataset [12] consists of real biographies collected from the web by filtering for lines that began with a name followed by the string "is a(n) (xxx) title," where title is an occupation from the BLS SOC system[1]. We model a scenario where experienced candidates apply for jobs in the same field. Each candidate is represented by non-sensitive features extracted from the text biography (term frequency of the occupation in the biography, length and number of words of the biography) and sensitive features: gender (provided by the dataset) and nationality (inferred from each candidate's name using the name2nat Python package [19]) under the assumption nationality is inferred from the name, as a recruiter might when reviewing a resume.

Data Pre-processing: Nationalities were grouped together by continent, but due to limited data in some of the intersectional groups, the nationalities were grouped to form an advantaged group (East-European and the West-European nationalities) and a disadvantaged group (African, Asian and Latin-American nationalities). The American nationality was discarded due to ambiguity between the inferred nationalities. The dataset doesn't include Spanish nationality, avoiding ambiguity with the Latin-American group. Train-test splits are stratified across intersectional groups, with five consistent splits per query using a 30% test set. **Relevance Judgements:** For each occupation, candidates are ranked by the cosine similarity between the word2vec [17] embedding of the occupation title and the text biography. Word2vec embeddings are known to perpetuate stereotypical associations [6, 14], simulating the pre-existing social bias in the data. The relevance judgements for training the model are assigned based on the cosine similarity with values between 1 and 500 (most relevant candidate), with scores below 0.4 considered negative samples for training. **Causal Estimation:** The causal model was estimated on the train set. Following prior research [25], the disadvantaged group, Female African-Asian-Latin, is chosen as the reference group. In experiments involving occupation modeling, psychologists are the reference occupation due to their balanced group distribution.

4. Results and Discussion

Fairness of a ranking is measured as the percentage of each sensitive group among the top 10. Our aim is to create a diverse ranked list of the candidates with respected to the intersectional sensitive groups by increasing the proportion of the underrepresented groups, without producing a swap between the underrepresented group and the over-represented group. If in one occupation females are underrepresented, we do not want to over-represent them. *Utility* of the ranking model is measured using Normalized Discounted Cumulative Gain across the top 10 (NDCG@10). We choose to evaluate at the top 10, as it is unlikely that the recruiter will scroll down or move to the next page to view more candidates [13, 18]. The results reported are an average over the five runs for each query.

RQ1: Do counterfactual representations lead to a diverse rank in a recruitment scenario? We report the results of Model 1 in Figure 2 - NoOccupation. We see that the counterfactual representations do not increase the proportion for any of the groups, except for the Male European group over the following occupations: filmmaker, journalist, software developer, surgeon, composer, painter and professor. This means that overall in the data the Male European group has lower scores due to the large number of candidates with lower scores in occupations where the Female African-Asian-Latin group are over-represented.

The results of estimating a causal model for each occupation (Figure 2 - SingleOccupation) show an increase in proportion across the sensitive groups for most occupations. The Female African-Asian-Latin group increased in proportion in the following occupations: physician, chiropractor, comedian, and software engineer. Interestingly, for comedian and software engineer, occupations fully over-represented by men in the top 10, the proportion increased for both female groups, African-Asian-Latin and European. The Male African-Asian-Latin group was increased in proportion for female dominated jobs, e.g., paralegal, teacher and yoga teacher, but also in occupations over-represented by Male Europeans, e.g., attorney and pastor. The Female European group was increased in male dominated jobs, e.g., poet, but also in jobs dominated by Female African-Asian-Latin, e.g., yoga teacher, model and dietitian.

Results (Figure 2 - ModeledOccupation) show that the changes in proportion are similar to the ones obtained by estimating a causal model for each occupation, with some minor fluctuations. For yoga teacher, the proportion of the ranking is more balanced using Model 2.

Table 1 shows how far the proportion of the groups is from achieving statistical parity, which is achieved when the probabilities of a favorable outcome are equal between the groups [20], meaning that all groups have equal proportion in top 10. Positive values indicate underrepresentation (0.25 means absence), approaching zero suggests a positive change, while negative values signify overrepresentation (-0.75 means exclusive presence). Distance from zero implies a negative change in proportion. The proportion increases often have a positive effect overall, meaning that the proportion was increased for the underrepresented groups, however, in some situations the increase in proportion of one group negatively affects another group (e.g., the occupation accountant is over-represented by men in the top 10, and regardless of this the Male African-Asian-Latin group is increased in proportion affecting negatively the proportion of the

[1] https://www.bls.gov/soc/

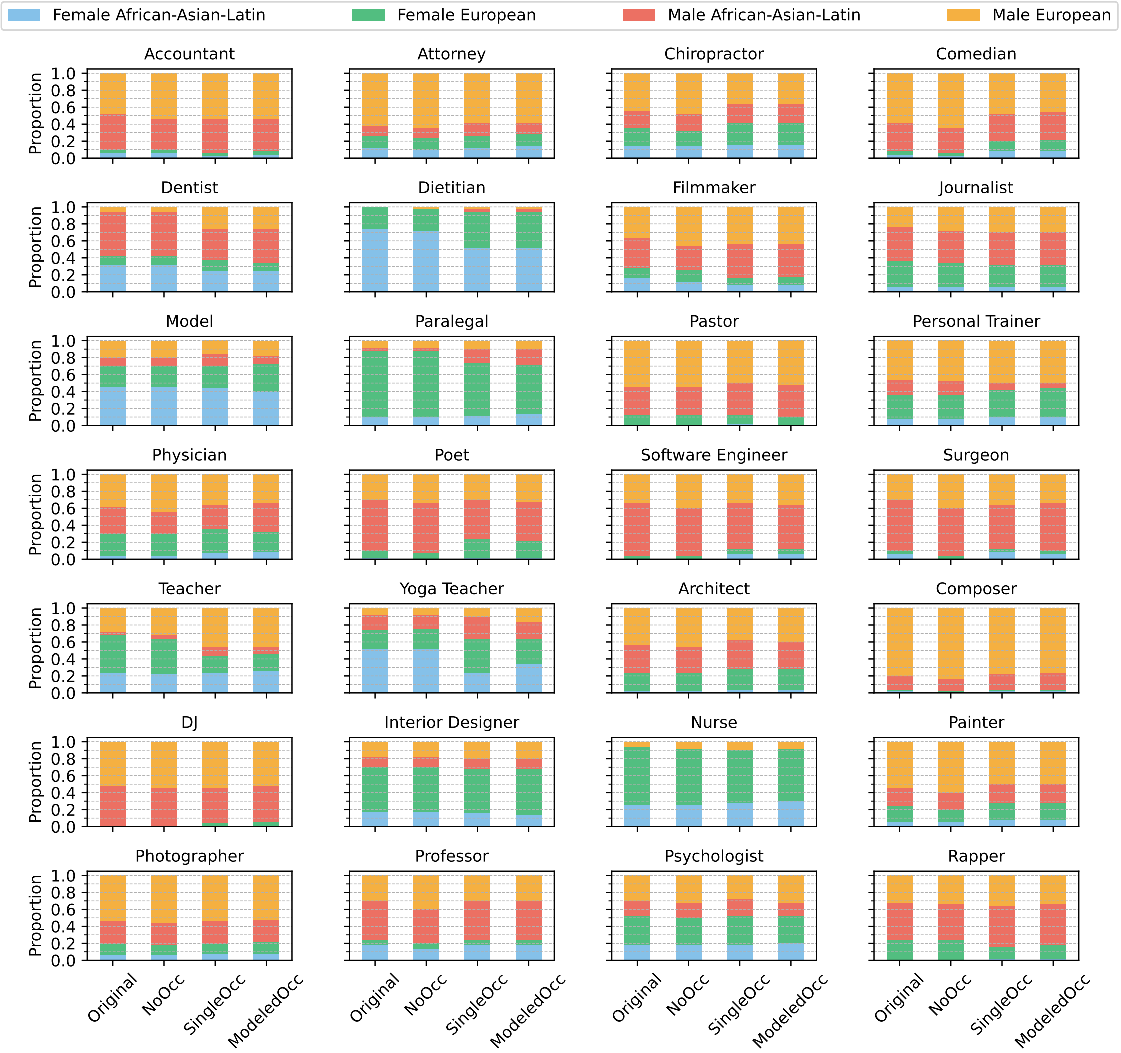


Figure 2: Distribution of the groups in the top 10. **Original** - original ranking of the candidates. **NoOcc** - Model 1 was applied over all occupations. **SingleOcc** - Model 1 was applied over each occupation. **ModeledOcc** - Model 2 was applied over all occupations.

females group which are underrepresented).

RQ2: How to model the bias encoded in each occupation? Prior work [25] has applied counterfactual intersectionality to datasets that contain a single ranked list. A recruitment scenario is more complex and challenging, as usually a recruiter needs a ranked list of candidates for each job opening, and the direction and degree of bias vary over occupations. For example, some occupations are female dominated, such as nurse, while others are male dominated, such as software developer [3]. Thus, for female dominated jobs one would want to increase the proportion of the males, while for male dominated jobs to increase the proportion of the females.

Applying the causal model over all occupations fails to capture the variations of the bias direction across the occupations. This results in creating counterfactual representations that change the observed representations based on an overall estimated bias in the data, which does not reflect the real bias associated with each occupation. Moreover, the bias estimates have very small values, resulting in small changes of the scores and of the rank.

One solution is to estimate a causal model for each occupation, or for groups of occupations with the same bias direction. We estimate a causal model for each occu-

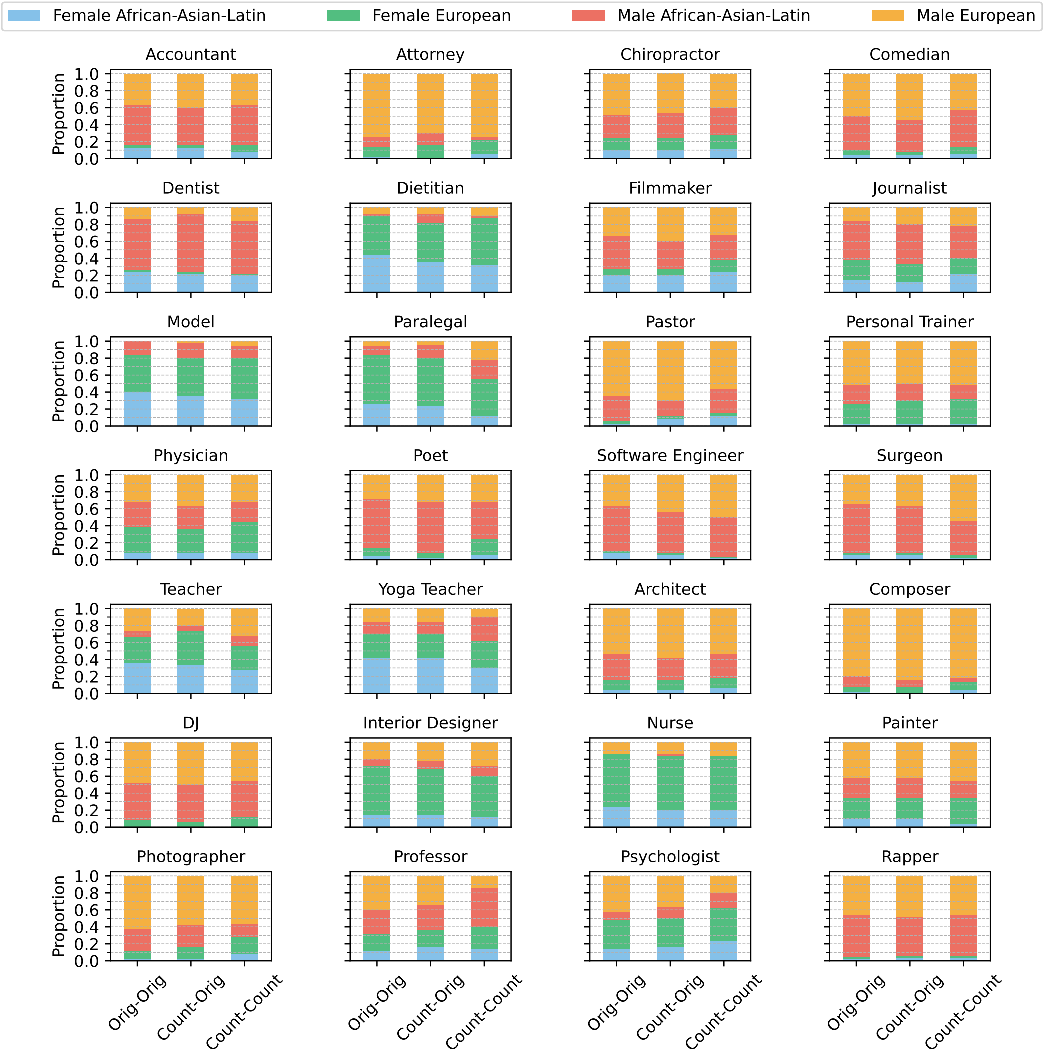


Figure 3: Distribution of the predicted groups in the top 10 by RankNet. **Orig-Orig** - the model was trained and tested on original representations. **Count-Count** - the model was trained and tested on the counterfactual representations computed using Model 1 on each occupation. **Count-Orig** - the model was trained on counterfactual representations and tested on the original representations.

pation without grouping, as there is no clear pattern for grouping the occupations due to the complexity added by the intersectional groups. Results (Figure 2 - SingleOccupation) show that estimating a causal model for each occupation captures the variations of bias in each occupation resulting in an increase in proportion across the sensitive groups for most occupations. Another solution to capture the variations of the direction of bias across occupations is to introduce a node in the causal graph representing the occupation. This indicates that the occupation has an influence on the bias present in the data, so that the bias estimate should be different for each occupation. Results (Figure 2 - ModeledOccupation) are similar to estimating a causal model for each occupation.

RQ3: Does training a ranking model on the counterfactual representations create a diverse ranking? Ideally, a model trained on fairer data would produce a fair ranking even without access to sensitive attributes. We use the Ranklib [10] implementation of RankNet [7], a pairwise learning to rank algorithm. Previous research [25] used ListNet [8], a listwise approach, but we observed RankNet has a better performance on the BIOS

Table 1
Statistical parity of the rankings, highlighting underrepresented and overrepresented group. **Orig** – original ranking of the candidates. **NoOcc** – Model 1 was applied over all occupations. **SOcc** – Model 1 was applied over each occupation. **MOcc** – Model 2 was applied over all occupations.

	Female African-Asian-Latin				Female European				Male African-Asian-Latin				Male European			
Occupation	Orig	NoOcc	SOcc	MOcc	Orig	NoOcc	SOcc	MOcc	Orig	NoOcc	SOcc	MOcc	Orig	NoOcc	SOcc	MOcc
Accountant	+0.19	+0.19	+0.23	+0.21	+0.21	+0.21	+0.21	+0.21	−0.17	−0.11	−0.15	−0.13	−0.23	−0.29	−0.29	−0.29
Architect	+0.23	+0.23	+0.21	+0.21	+0.03	+0.03	+0.01	+0.01	−0.07	−0.05	−0.09	−0.07	−0.19	−0.21	−0.13	−0.15
Attorney	+0.13	+0.15	+0.13	+0.11	+0.11	+0.11	+0.11	+0.11	+0.13	+0.13	+0.09	+0.11	−0.37	−0.39	−0.33	−0.33
Chiropractor	+0.11	+0.11	+0.09	+0.09	+0.03	+0.07	−0.01	−0.01	+0.05	+0.05	+0.03	+0.03	−0.19	−0.23	−0.11	−0.11
Comedian	+0.21	+0.23	+0.17	+0.17	+0.21	+0.21	+0.13	+0.11	−0.09	−0.05	−0.07	−0.07	−0.33	−0.39	−0.23	−0.21
Composer	+0.23	+0.25	+0.23	+0.23	+0.23	+0.23	+0.23	+0.23	+0.09	+0.11	+0.07	+0.05	−0.55	−0.59	−0.53	−0.51
Dentist	−0.07	−0.07	+0.01	+0.01	+0.15	+0.15	+0.11	+0.15	−0.27	−0.27	−0.11	−0.15	+0.19	+0.19	−0.01	−0.01
Dietitian	−0.49	−0.47	−0.27	−0.27	−0.01	−0.01	−0.17	−0.17	+0.25	+0.25	+0.21	+0.21	+0.25	+0.23	+0.23	+0.23
Dj	+0.25	+0.25	+0.25	+0.25	+0.25	+0.25	+0.21	+0.19	−0.23	−0.21	−0.17	−0.17	−0.27	−0.29	−0.29	−0.27
Filmmaker	+0.09	+0.13	+0.17	+0.17	+0.13	+0.11	+0.17	+0.15	−0.11	−0.03	−0.15	−0.13	−0.11	−0.21	−0.19	−0.19
Interior Designer	+0.07	+0.07	+0.09	+0.11	−0.27	−0.27	−0.27	−0.29	+0.13	+0.13	+0.13	+0.13	+0.07	+0.07	+0.05	+0.05
Journalist	+0.19	+0.19	+0.19	+0.19	−0.05	−0.03	−0.01	−0.01	−0.15	−0.13	−0.13	−0.13	+0.01	−0.03	−0.05	−0.05
Model	−0.21	−0.21	−0.19	−0.15	+0.01	+0.01	−0.01	−0.07	+0.15	+0.15	+0.11	+0.15	+0.05	+0.05	+0.09	+0.07
Nurse	−0.01	−0.01	−0.03	−0.05	−0.43	−0.41	−0.37	−0.37	+0.25	+0.25	+0.25	+0.25	+0.19	+0.17	+0.15	+0.17
Painter	+0.19	+0.19	+0.17	+0.17	+0.07	+0.11	+0.05	+0.05	+0.03	+0.05	+0.03	+0.03	−0.29	−0.35	−0.25	−0.25
Paralegal	+0.15	+0.15	+0.13	+0.11	−0.53	−0.53	−0.37	−0.33	+0.21	+0.21	+0.09	+0.07	+0.17	+0.17	+0.15	+0.15
Pastor	+0.25	+0.25	+0.23	+0.25	+0.13	+0.13	+0.15	+0.15	−0.09	−0.09	−0.13	−0.13	−0.29	−0.29	−0.25	−0.27
Personal Trainer	+0.17	+0.17	+0.15	+0.15	−0.03	−0.03	−0.07	−0.09	+0.07	+0.09	+0.17	+0.19	−0.21	−0.23	−0.25	−0.25
Photographer	+0.19	+0.19	+0.17	+0.17	+0.11	+0.13	+0.13	+0.11	−0.01	−0.01	−0.01	−0.01	−0.29	−0.31	−0.29	−0.27
Physician	+0.21	+0.21	+0.17	+0.17	−0.01	−0.01	−0.03	+0.01	−0.07	−0.01	−0.03	−0.09	−0.13	−0.19	−0.11	−0.09
Poet	+0.23	+0.25	+0.23	+0.23	+0.17	+0.17	+0.03	+0.05	−0.35	−0.33	−0.21	−0.21	−0.05	−0.09	−0.05	−0.07
Professor	+0.07	+0.11	+0.07	+0.07	+0.19	+0.19	+0.19	+0.19	−0.21	−0.15	−0.21	−0.21	−0.05	−0.15	−0.05	−0.05
Psychologist	+0.07	+0.07	+0.07	+0.05	−0.09	−0.07	−0.09	−0.07	+0.07	+0.07	+0.05	+0.09	−0.05	−0.07	−0.03	−0.07
Rapper	+0.25	+0.25	+0.23	+0.23	+0.01	+0.01	+0.11	+0.09	−0.19	−0.17	−0.23	−0.23	−0.07	−0.09	−0.11	−0.09
Software Engineer	+0.25	+0.25	+0.19	+0.19	+0.21	+0.21	+0.19	+0.19	−0.37	−0.31	−0.29	−0.27	−0.09	−0.15	−0.09	−0.11
Surgeon	+0.19	+0.25	+0.17	+0.19	+0.21	+0.21	+0.21	+0.21	−0.35	−0.31	−0.27	−0.31	−0.05	−0.15	−0.11	−0.09
Teacher	+0.01	+0.03	+0.01	−0.01	−0.19	−0.17	+0.05	+0.05	+0.21	+0.21	+0.15	+0.17	−0.03	−0.07	−0.21	−0.21
Yoga Teacher	−0.27	−0.27	+0.01	−0.09	+0.03	+0.01	−0.15	−0.05	+0.07	+0.09	−0.01	+0.05	+0.17	+0.17	+0.15	+0.09

dataset. RankNet obtained an NDCG@10 of 48% as opposed to ListNet, which obtained 35%. The performance of RankNet when trained on counterfactual representations did not decrease. When tested on counterfactual representations it obtained 57%, compared to 48% on the original representations.

Figure 3 shows the distributions of the groups across occupations. The counterfactual representations are computed using Model 1, but similar results were observed using Model 2. Ideally we would want to see that RankNet trained on counterfactual representations (Count-Count) creates a more diverse ranking in the top 10, and that when tested on the original representations this still holds (Count-Orig). Results (Figure 3, Count-Orig) show that the occupations for which the proportion is positively increased for at least one group are: attorney, comedian, journalist, poet, software engineer, yoga teacher, dj, nurse, painter, photographer and professor. Overall it seems it mostly increases the proportion for one of the under-represented groups. For example, for professor, it does not increase the proportion for the Female European group, but it increases it for the Female African-Asian-Latin group. There are occupations where we can observe there is change in proportion but it affects negatively the balance in proportion, increasing the over-represented groups.

Table 2 shows how far the proportion of the groups is from achieving statistical parity. We confirm the results of previous research which showed that even when not knowing the sensitive attributes, at inference time, some increase in proportion can be observed. However, we observed this only for some occupations, and further investigation over when this holds and when it does not, should be performed to be able to safely use a model to predict a fairer ranking of the candidates without having access to the sensitive attributes.

5. Conclusion

In this work we consider the applicability of existing fairness methods to recruitment. Legal requirements make many approaches difficult to use in practice, as access to special sensitive attributes is limited. We argue that pre-processing methods are well-suited to creating a diverse ranking of candidates, and we propose to use a counterfactual method to create fairer representations for candidates. The counterfactual method additionally makes score adjustments explicit, which can help satisfy transparency requirements. To apply this approach to the recruitment scenario, it is necessary to either estimate a causal model for each occupation or group of occupations, or to add a node representing the occupation to the causal graph.

This approach was evaluated on the BIOS dataset,

Table 2

Statistical parity of the predicted groups, highlighting underrepresented and overrepresented group, in the top 10 by RankNet. **OO** – the model was trained and tested on original representations. **CC** – the model was trained and tested on the counterfactual representations. **CO** – the model was trained on counterfactual representations and tested on the original representations.

	Female African-Asian-Latin			Female European			Male African-Asian-Latin			Male European		
Occupation	**OO**	**CC**	**CO**	**OO**	**CC**	**CO**	**OO**	**CC**	**CO**	**OO**	**CC**	**CO**
Accountant	+0.13	+0.17	+0.13	+0.21	+0.17	+0.21	−0.23	−0.23	−0.19	−0.11	−0.11	−0.15
Architect	+0.21	+0.19	+0.21	+0.13	+0.13	+0.13	−0.05	−0.03	−0.01	−0.29	−0.29	−0.33
Attorney	+0.23	+0.19	+0.25	+0.13	+0.09	+0.09	+0.13	+0.21	+0.11	−0.49	−0.49	−0.45
Chiropractor	+0.15	+0.13	+0.15	+0.11	+0.09	+0.11	−0.03	−0.07	−0.05	−0.23	−0.15	−0.21
Comedian	+0.21	+0.19	+0.21	+0.19	+0.17	+0.21	−0.15	−0.19	−0.13	−0.25	−0.17	−0.29
Composer	+0.23	+0.21	+0.25	+0.19	+0.15	+0.17	+0.13	+0.21	+0.17	−0.55	−0.57	−0.59
Dentist	+0.01	+0.05	+0.03	+0.23	+0.23	+0.23	−0.35	−0.37	−0.43	+0.11	+0.09	+0.17
Dietitian	−0.19	−0.07	−0.11	−0.21	−0.31	−0.21	+0.23	+0.23	+0.15	+0.17	+0.15	+0.17
Dj	+0.25	+0.25	+0.25	+0.17	+0.13	+0.19	−0.19	−0.17	−0.19	−0.23	−0.21	−0.25
Filmmaker	+0.05	+0.01	+0.05	+0.17	+0.11	+0.17	−0.13	−0.05	−0.07	−0.09	−0.07	−0.15
Interior Designer	+0.11	+0.13	+0.11	−0.33	−0.23	−0.29	+0.17	+0.13	+0.15	+0.05	−0.03	+0.03
Journalist	+0.11	+0.03	+0.13	+0.01	+0.07	+0.03	−0.21	−0.13	−0.21	+0.09	+0.03	+0.05
Model	−0.15	−0.07	−0.11	−0.19	−0.23	−0.19	+0.09	+0.11	+0.07	+0.25	+0.19	+0.23
Nurse	+0.01	+0.05	+0.05	−0.37	−0.39	−0.39	+0.25	+0.25	+0.23	+0.11	+0.09	+0.11
Painter	+0.15	+0.21	+0.15	+0.01	−0.05	+0.01	+0.01	+0.05	+0.01	−0.17	−0.21	−0.17
Paralegal	−0.01	+0.13	+0.01	−0.33	−0.19	−0.31	+0.15	+0.03	+0.09	+0.19	+0.03	+0.21
Pastor	+0.23	+0.13	+0.17	+0.21	+0.21	+0.21	−0.05	−0.03	+0.07	−0.39	−0.31	−0.45
Personal Trainer	+0.23	+0.23	+0.23	+0.01	−0.05	−0.03	+0.03	+0.09	+0.05	−0.27	−0.27	−0.25
Photographer	+0.23	+0.17	+0.23	+0.15	+0.05	+0.11	−0.01	+0.09	−0.01	−0.37	−0.31	−0.33
Physician	+0.17	+0.17	+0.17	−0.05	−0.11	−0.03	−0.05	+0.01	−0.03	−0.07	−0.07	−0.11
Poet	+0.21	+0.19	+0.23	+0.15	+0.07	+0.19	−0.33	−0.19	−0.35	−0.03	−0.07	−0.07
Professor	+0.13	+0.11	+0.09	+0.05	−0.01	+0.05	−0.03	−0.21	−0.05	−0.15	+0.11	−0.09
Psychologist	+0.11	+0.01	+0.09	−0.09	−0.13	−0.09	+0.15	+0.07	+0.11	−0.17	+0.05	−0.11
Rapper	+0.23	+0.21	+0.21	+0.23	+0.23	+0.23	−0.25	−0.23	−0.21	−0.21	−0.21	−0.23
Software Engineer	+0.17	+0.23	+0.19	+0.23	+0.23	+0.23	−0.29	−0.21	−0.23	−0.11	−0.25	−0.19
Surgeon	+0.19	+0.23	+0.19	+0.23	+0.21	+0.23	−0.33	−0.15	−0.31	−0.09	−0.29	−0.11
Teacher	−0.11	−0.03	−0.09	−0.05	−0.03	−0.15	+0.17	+0.13	+0.19	−0.01	−0.07	+0.05
Yoga Teacher	−0.17	−0.05	−0.17	−0.03	−0.07	−0.03	+0.11	−0.03	+0.11	+0.09	+0.15	+0.09

where we show that a model trained on the counterfactual representations can create a more diverse ranking, without having access to sensitive attributes at inference time. These results confirm that modeling the occupation is important in a recruitment scenario where different occupations may be associated with different biases. While the BIOS dataset is a reasonable proxy, the data does not come from a real recruitment scenario and may differ in several ways (e.g., distributions of people and bias may differ from a real scenario, the dataset contains less detailed information about education and work experience, and including artistic occupations might be unusual in a hiring process).

Future work could investigate under what conditions the counterfactual representations lead to an increase in diversity that creates a balanced distribution of intersectional groups in top-ranked positions, given real features for candidates such as education and work experience.

Acknowledgments

This research was (partially) funded by the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union’s Horizon Europe research and innovation program under grant agreement No. 101070212, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, and project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] General data protection regulation (GDPR). https://gdpr-info.eu/. Accessed: 2023-07-31.

[2] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–30, 2019.

[3] F. Bettio, A. Verashchagina, I. Mairhuber, D. Meulders, I. Beleva, A. Panayiotou, A. Křižkova, and R. Emerek. *Gender segregation in the labour market: Root causes, implications and policy responses in the EU*. European Commission, 2009.

[4] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019.

[5] P. Bisschop, B. ter Weel, and J. Zwetsloot. Ethnic employment gaps of graduates in the Netherlands. *De Economist*, 168(4):577–598, 2020.

[6] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29, 2016.

[7] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, 2005.

[8] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, 2007.

[9] G. Ciminelli, C. Schwellnus, and B. Stadler. Sticky floors or glass ceilings? The role of human capital, working time flexibility and discrimination in the gender wage gap. Technical Report 1668, OECD Publishing, 2021.

[10] V. Dang and C. Ascent. The Lemur project. http://lemurproject.org, 2023.

[11] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[12] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

[13] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 478–479, 2004.

[14] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[15] P. Lahoti, K. P. Gummadi, and G. Weikum. iFair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th international Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

[16] E. Matteazzi, A. Pailhé, and A. Solaz. Part-time employment, the gender wage gap and the role of wage-setting institutions: Evidence from 11 European countries. *European Journal of Industrial Relations*, 24(3):221–241, 2018.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013. doi: 10.48550/arXiv.1301.3781.

[18] M. O'Brien and M. T. Keane. Modeling result-list searching in the world wide web: The role of relevance topologies and trust bias. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, volume 28, pages 1881–1886. Citeseer, 2006.

[19] K. Park. name2nat: A Python package for nationality prediction from a name. https://github.com/Kyubyong/name2nat, 2020.

[20] E. Pitoura, K. Stefanidis, and G. Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28, 2022.

[21] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018.

[22] L. Thijssen, B. Lancee, S. Veit, and R. Yemane. Discrimination against Turkish minorities in Germany and the Netherlands: Field experimental evidence on the effect of diagnostic information on labour market outcomes. *Journal of Ethnic and Migration Studies*, 47(6):1222–1239, 2021.

[23] M. van Bekkum and F. Borgesius Zuiderveen. Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review*, 48:105770, 2023.

[24] K. Yang, V. Gkatzelis, and J. Stoyanovich. Balanced ranking with diversity constraints. *arXiv preprint arXiv:1906.01747*, 2019.

[25] K. Yang, J. R. Loftus, and J. Stoyanovich. Causal intersectionality for fair ranking. *arXiv preprint*

arXiv:2006.08688, 2020.

[26] Q. Yu and B. Li. mma: An R package for mediation analysis with multiple mediators. *Journal of Open Research Software*, 5(1), 2017.

[27] M. Zehlike and C. Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of the Web Conference 2020*, pages 2849–2855, 2020.

[28] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

[29] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, Part i: Score-based ranking. *ACM Computing Surveys*, 55(6):1–36, 2022.

[30] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333. PMLR, 2013.