findhr UNIVERSITY OF AMSTERDAM

Counterfactual Representations for Intersectional Fair Ranking in Recruitment

Clara Rus, Maarten de Rijke and Andrew Yates

Problem

Amazon scraps secret AI recruiting tool that showed bias against women

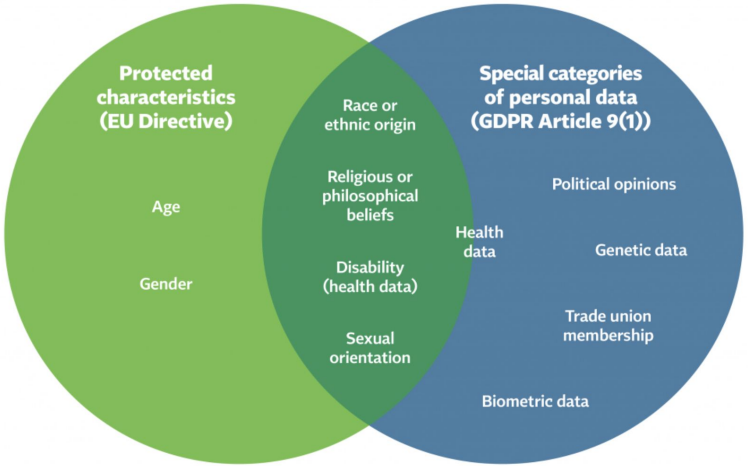
Facebook's ad delivery system still has gender bias, new study finds

/ Women are excluded from seeing some job listings 'beyond what can be legally justified'

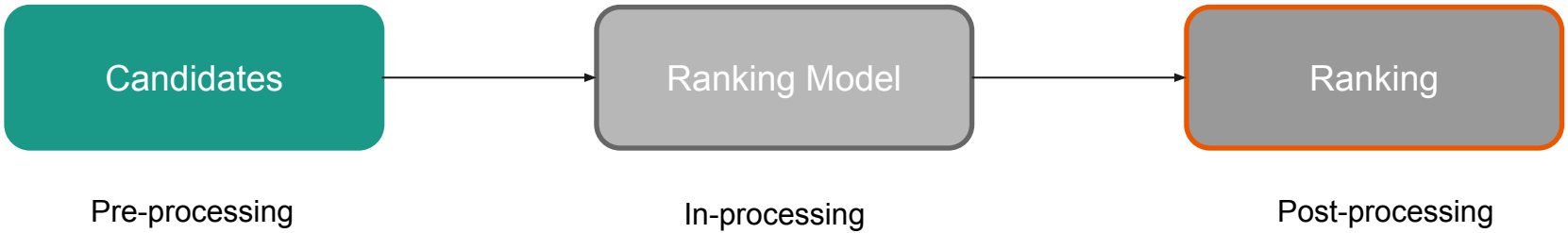
Fairness Interventions

Art. 9 GDPR

1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

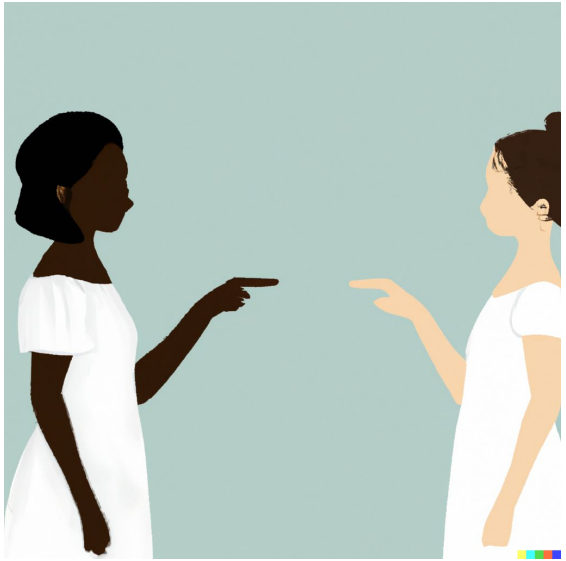


Fairness Interventions



Counterfactual Intersectionality for Fair Ranking

"What would this person's data look like if they had (or had not) been a Black woman (for example)?"



Intersectionality

Achieving fairness for each gender subgroup (e.g., men and women) and for each racial subgroup (e.g., black and white) does not guarantee a fair representation for a subgroup defined by the intersection of both attributes (e.g., black women)

→ **candidates belonging to multiple sensitive groups are subject to more discrimination**

Method

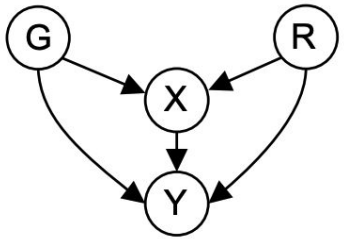
Step 1: Construct the causal model describing the data

Step 2: Estimate the total effect of the sensitive attributes (G-gender, R-race) on the score (Y) and on the features (X)

Step 3: Compute the counterfactual scores and features

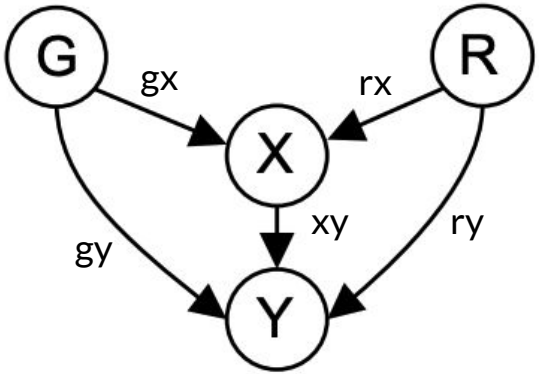
Step 4: Rank candidates according to the counterfactual score

→ ranking is fair with respect to race, gender, and the intersectional subgroups of these categories



Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2021. Causal intersectionality and fair ranking. In Symposium on Foundations of Responsible Computing (FORC). https://doi.org/10.4230/LIPIcs.FORC.2021.7

Method



Total Effect = direct effect of sensitive attributes (ry and gy) + indirect effect

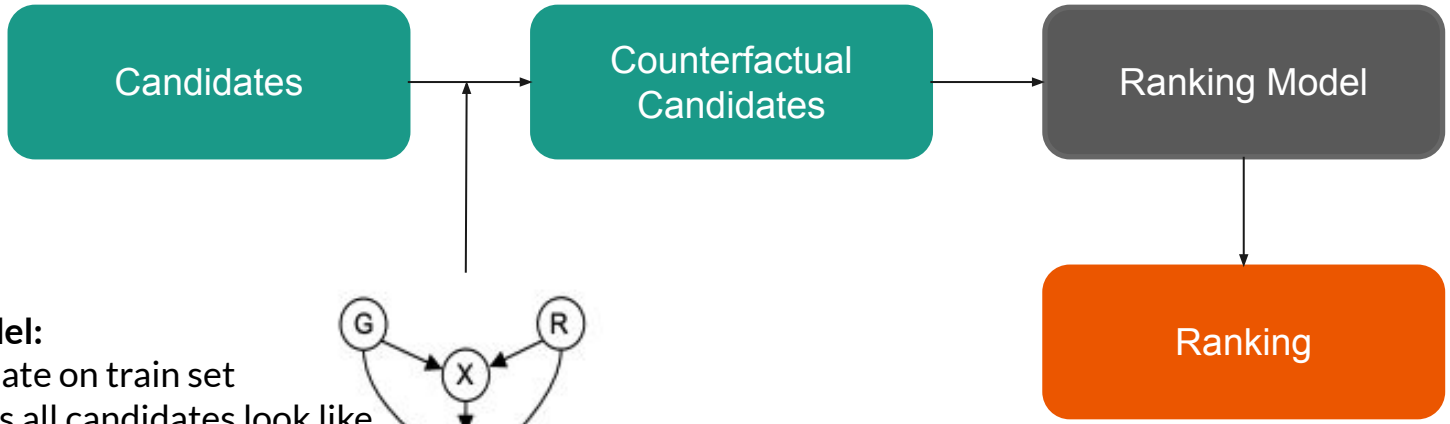
Indirect Effect = the effect of sensitive attributes that is mediated by the features (gx–xy and rx–ry)

Y counterfactual = Y observed + difference in Total Effect of the actual group and the control group

Experimental Setup

Task: given an occupation rank candidates

RankNet



Causal model:

- estimate on train set
- makes all candidates look like Female African-Asian-Latin

Evaluation: Proportion@10 = percentage of each group in top-10

Experimental Setup - Data

Dataset: Bios Bias - biographies of real people

Items: candidates - text bio

- length of bio
- number of words
- TF of occupation in the bio

Score: cosine similarity between bio and occupation

Sensitive attributes:

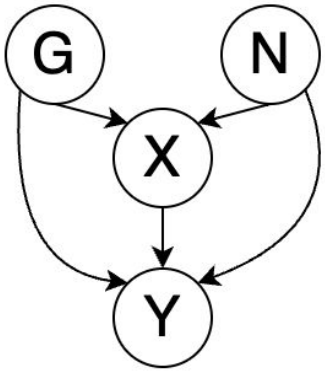
- gender (Male, Female)
- nationality (European, African-Asian-Latin)
  → inferred from name

Man is to Computer Programmer as Woman is to Homemaker

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Adam Kalai. Bias in Bios: A Case Study of Semantic Representation Bias in a High Stakes Setting. Proceedings of FAT*, 2019.

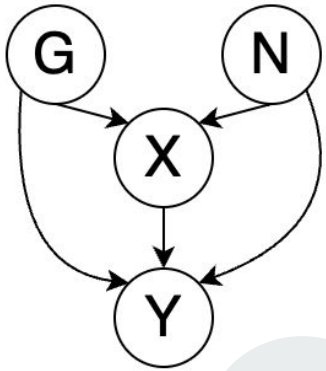
How to model the bias encoded in each occupation?

No Occupation Modelled



Model

Modelled Single Occupation

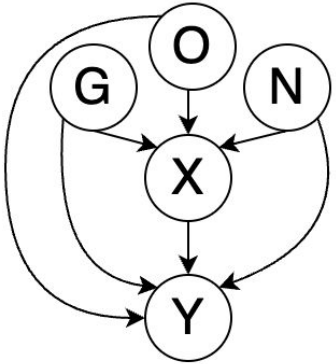


Model
Nurse

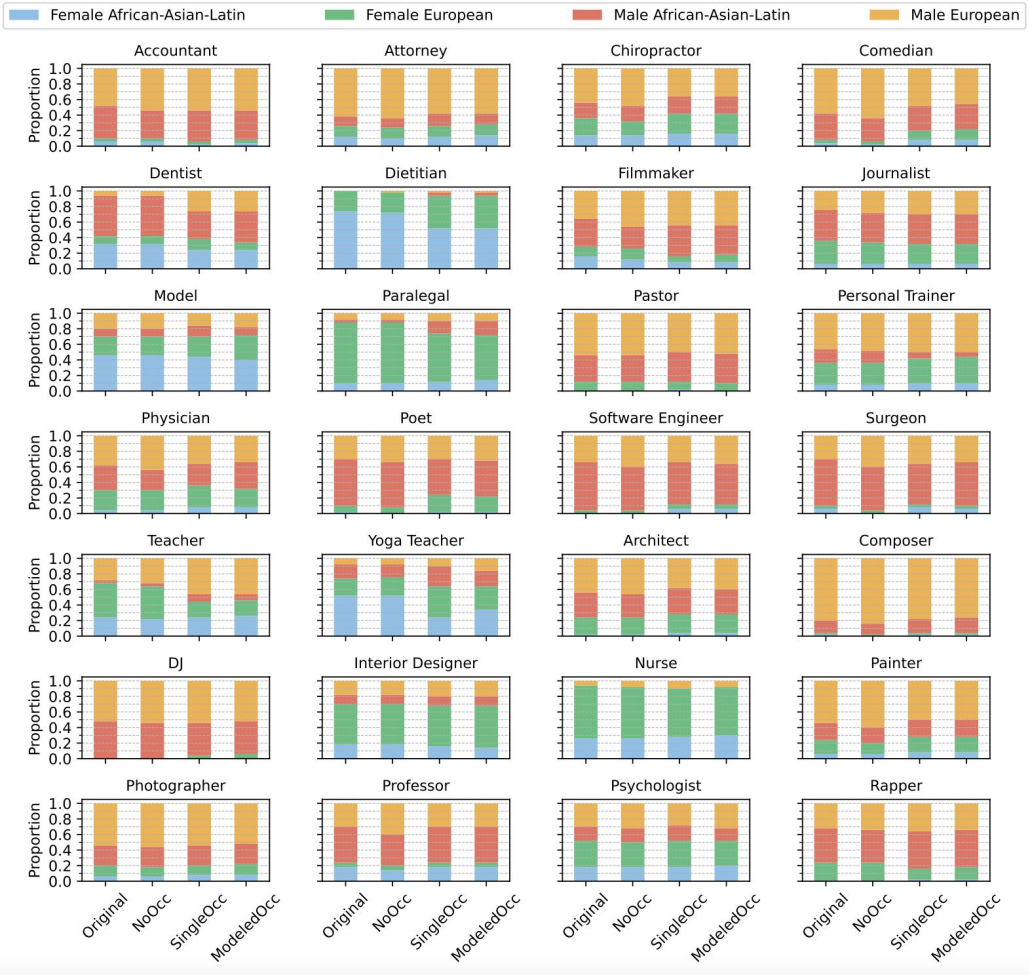
Model
Accountant

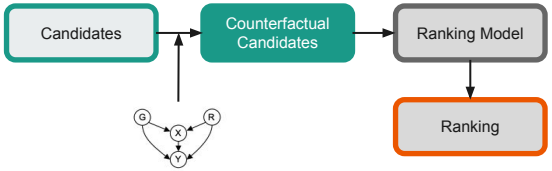
...

Modelled Occupation



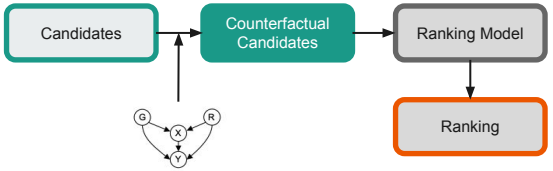
Model

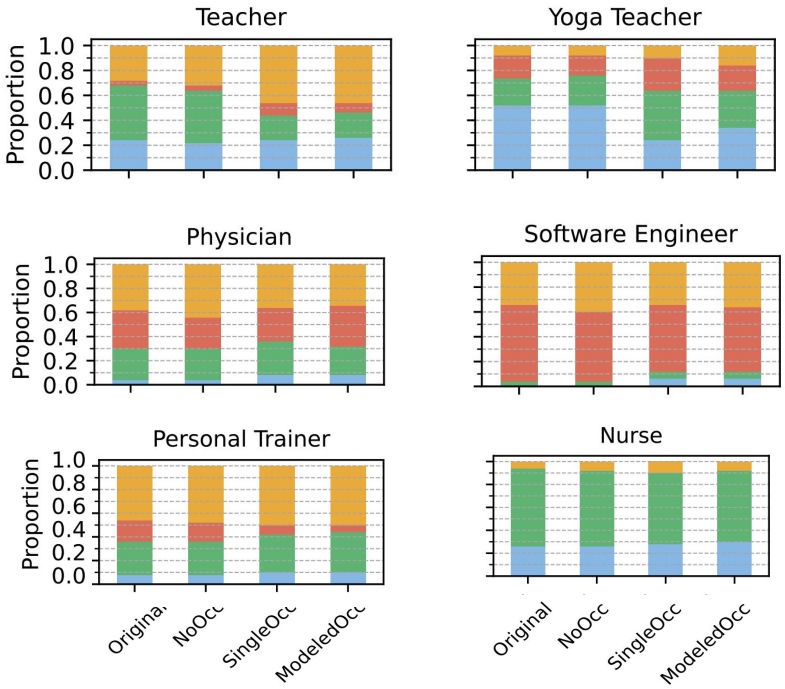




Do counterfactual representations lead to a diverse rank in a recruitment scenario?

Counterfactual Representations



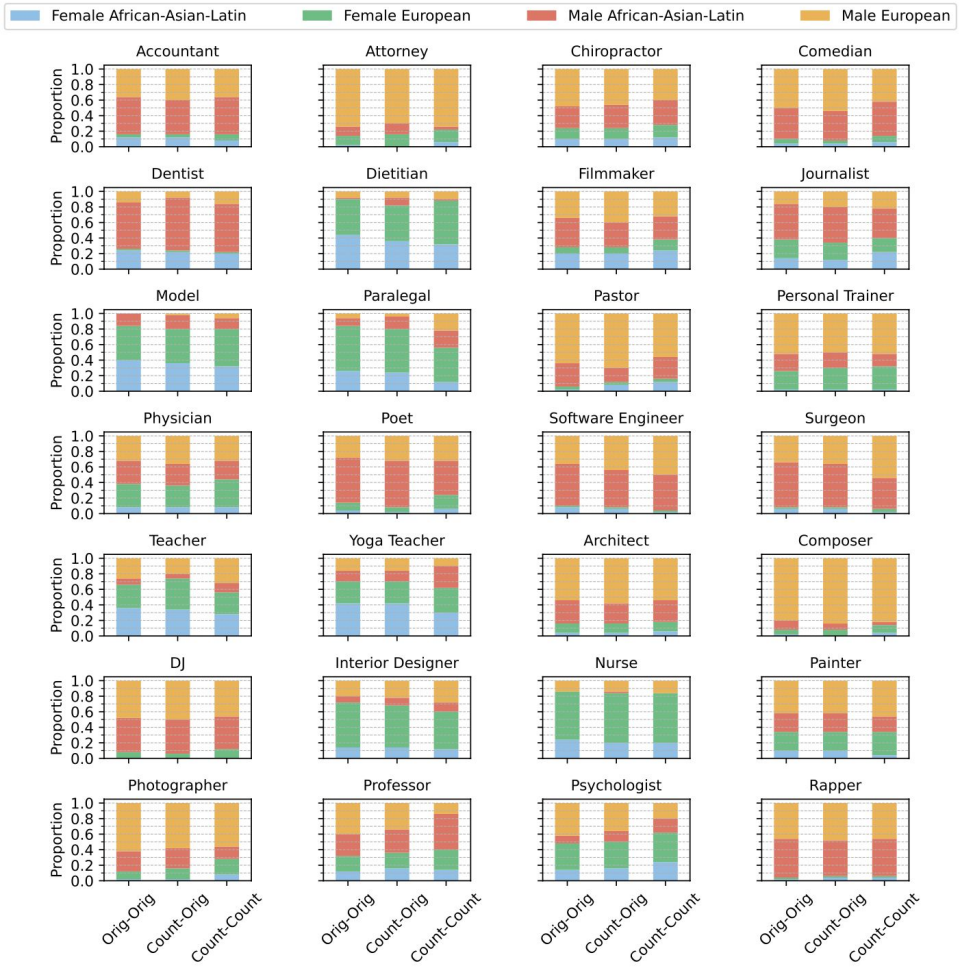


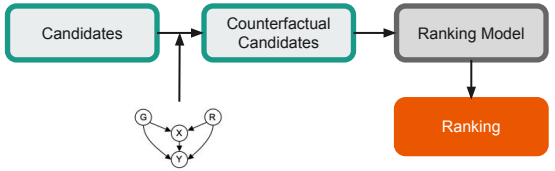
Female Dominated Jobs

Male Dominated Jobs

Not a positive change

- Female African-Asian-Latin
- Female European
- Male African-Asian-Latin
- Male European





Does training a ranking model on the counterfactual representations create a diverse ranking?

Conclusions

Legal requirements make many approaches difficult to use in practice ⟶ pre-processing techniques

Counterfactual Method
⟶ satisfy transparency requirements
⟶ provides an intersectional framework
⟶ showed promising results

Model the occupation in some way ⟶ bias direction varies

Future Work

⟶Under what conditions the counterfactual representations lead to an increase in diversity given real features for candidates such as education and work experience.

Conclusions

Legal requirements make many approaches difficult to use in practice ⟶ pre-processing techniques

Counterfactual Method
⟶ satisfy transparency requirements
⟶ provides an intersectional framework
⟶ showed promising results

Model the occupation in some way ⟶ bias direction varies

Future Work

⟶Under what conditions the counterfactual representations lead to an increase in diversity given real features for candidates such as education and work experience.

Thank you!

Data Donation Campaign

DATA DONATION CAMPAIGN

Help us make hiring fair

Help us prevent discrimination in job application processes. It has to stop that people are singled out because of their gender identity, alleged “ethnicity” or sexual orientation! **What can you do?** Donate your (anonymized) CV, answer (optional) questions about sensitive data.

You have questions? Have a look at our FAQs and information sheet 👇

DONATE YOUR CV

⏳This survey takes 10 minutes.

📄Upload your anonymized CV.

❓Answer optional questions about sensitive data.