The University of Texas at Austin
McCombs School of Business

Social Norm Bias: Residual Harms of Fairness-Aware Algorithms

Maria De-Arteaga, PhD

Assistant Professor
Information, Risk and Operations Management Department
University of Texas at Austin

Algorithmic group fairness

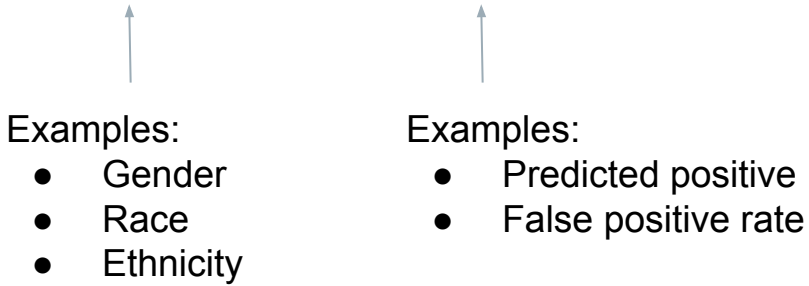
- Measures disparities across **sensitive groups** on a **measure of interest**.

Examples:
- Gender
- Race
- Ethnicity

Examples:
- Predicted positive
- False positive rate

Algorithmic group fairness

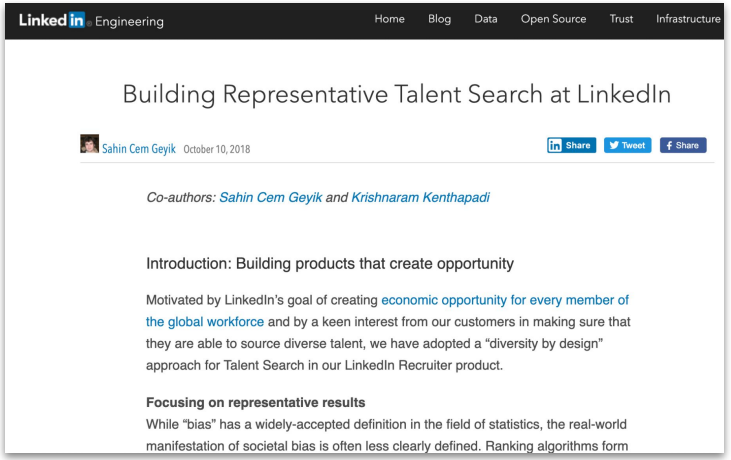
- Measures disparities across **sensitive groups** on a **measure of interest**.



- Group fairness constraints enforced via different strategies:
  - pre-processing: transformation of input data
  - in-processing: constraints in optimization objective
  - post-processing: transformation of output scores

Algorithmic group fairness

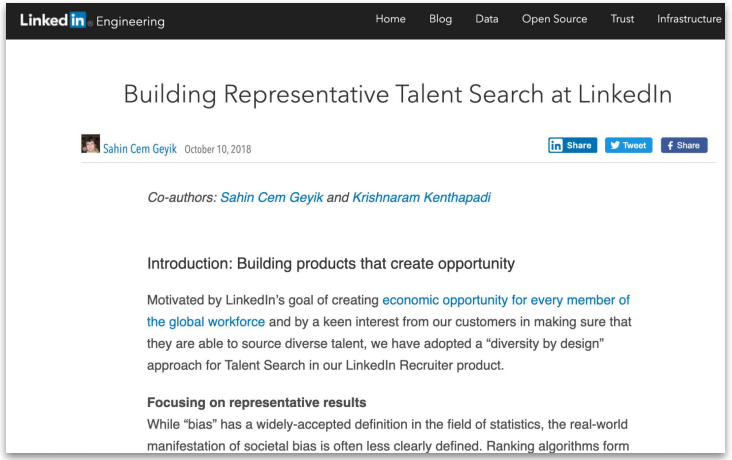
- Very popular due to ease of operationalization, increasingly implemented in ML deployment.
  - Especially true of post-processing (no cost of retraining!)
- Really great that algorithmic fairness research is having an impact! Are we done? Is the problem fixed?



Post-processing

Algorithmic group fairness

- Very popular due to ease of operationalization, increasingly implemented in ML deployment.
  - Especially true of post-processing (no cost of retraining!)
- Really great that algorithmic fairness research is having an impact! Are we done? Is the problem fixed?
  - Not really! Risk of reductive definitions and invisibilized harms.



What's Sex Got to Do With Fair Machine Learning?[1]

Lily Hu[2] & Issa Kohler-Hausmann[3]

Algorithmic Fairness from a Non-ideal Perspective

Sina Fazelpour & Zachary C. Lipton

What are some of the invisibilized harms?

What are some of the invisibilized harms?

Social Norm Bias: Residual Harms of Fairness-Aware Algorithms

Myra Cheng, Maria De-Arteaga, Lester Mackey, Adam Tauman Kalai

ICML Machine Learning for Data Workshop, ICML Socially Responsible Machine Learning Workshop



What are some of the invisibilized harms?

Social Norm Bias: Residual Harms of Fairness-Aware Algorithms

Myra Cheng, Maria De-Arteaga, Lester Mackey, Adam Tauman Kalai

ICML Machine Learning for Data Workshop, ICML Socially Responsible Machine Learning Workshop



Social Norm Bias (SNoB): The associations between an **algorithm's predictions** and **adherence to social norms**

Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder

Madeline E. Heilman*
New York University

How Women Engineers Do and Undo Gender: Consequences for Gender Equality

Abigail Powell, Barbara Bagilhole, Andrew Dainty

Gender stereotypes and workplace bias

Madeline E. Heilman

Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder

Madeline E. Heilman*
New York University

How Women Engineers Do and Undo Gender: Consequences for Gender Equality

Abigail Powell, Barbara Bagilhole, Andrew Dainty

Gender stereotypes and workplace bias

Madeline E. Heilman

Do algorithms exhibit social norm bias?

Is this bias mitigated by group fairness approaches?

In other words, am I more likely to benefit from group fairness approaches if I sound/act "like a man"?

Experiments using task/dataset *Bias in Bios*

she is a fifth year phd student in the joint machine learning and public policy program at carnegie mellon university <unk> s machine learning department and heinz college . she is co-advised by prof. artur <unk> and prof. alexandra chouldechova , and she is part of the auton lab . currently , her main focus is algorithmic fairness , studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support . she is passionate about developing novel machine learning algorithms that are motivated by existing policy problems , and understanding how machine learning can better help us overcome important societal challenges . prior to graduate school she received her b.sc . in mathematics from universidad nacional de colombia and worked as a journalist for one of colombia <unk> s main news magazine , semana . she is the recipient of a microsoft

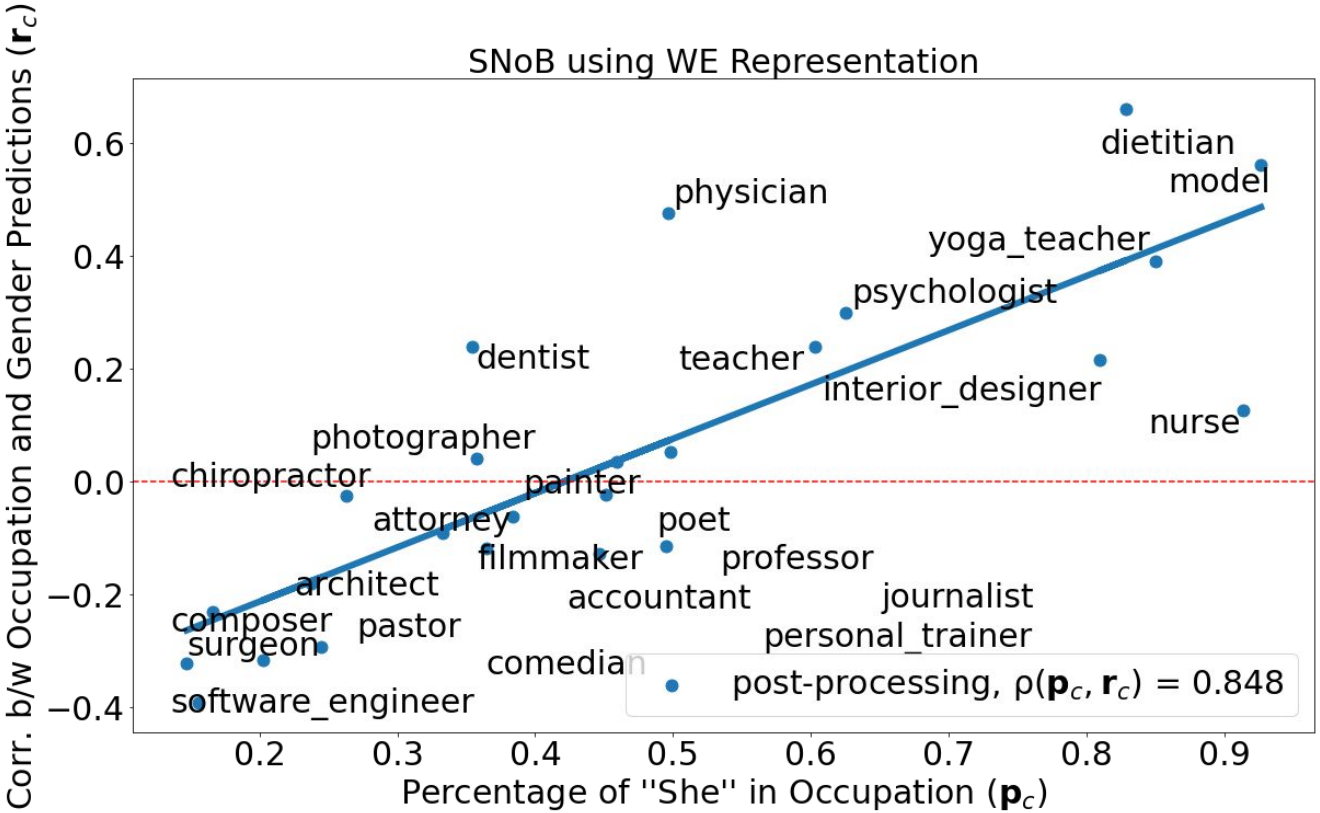
teacher

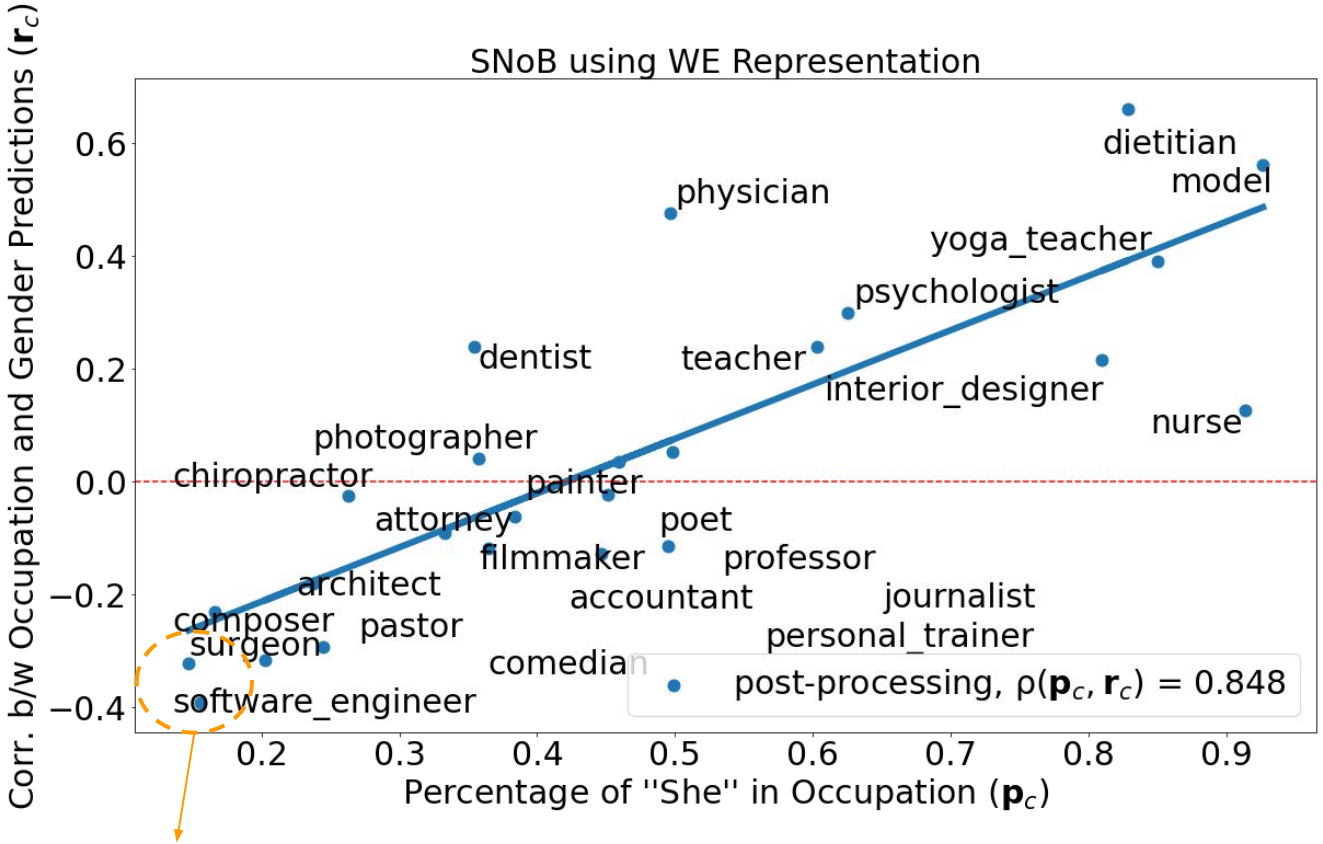
Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting (FAccT 2019)
M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. Kalai

Measuring Inferred Social Norms

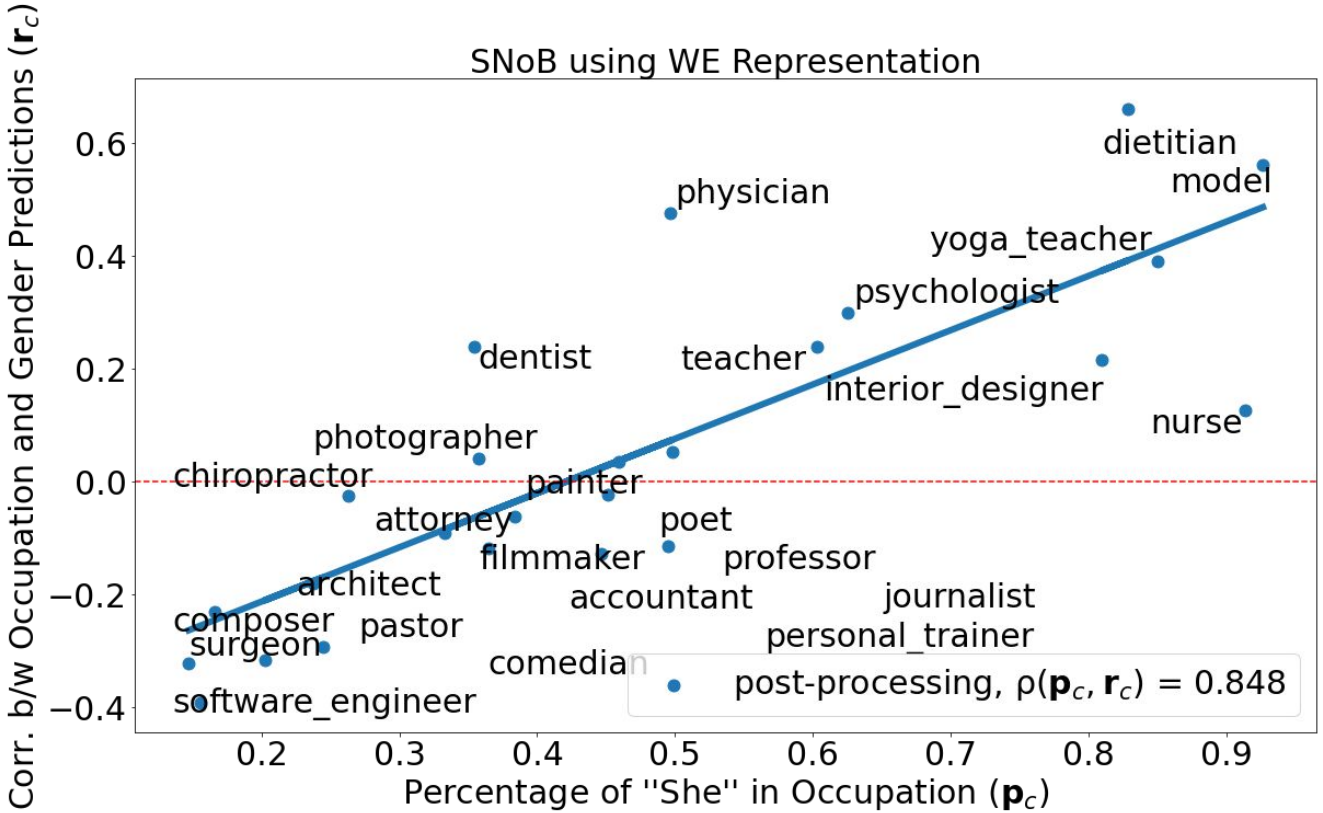
- Trained a classifier **G(x)** to predict group membership (e.g. “he”/“she”) from bio *x*
  - Gender-balanced within occupation
- Use probability of “he”/“she” as measure of aligning to inferred masculine/feminine gender norms
  - Does the biography align with what the algorithm associates as masculine/feminine?
  - Validation to ensure associations learned by algorithms align with human’s gender norm associations

Cryan, Jenna, et al. "Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods." ACM CHI, 2020.

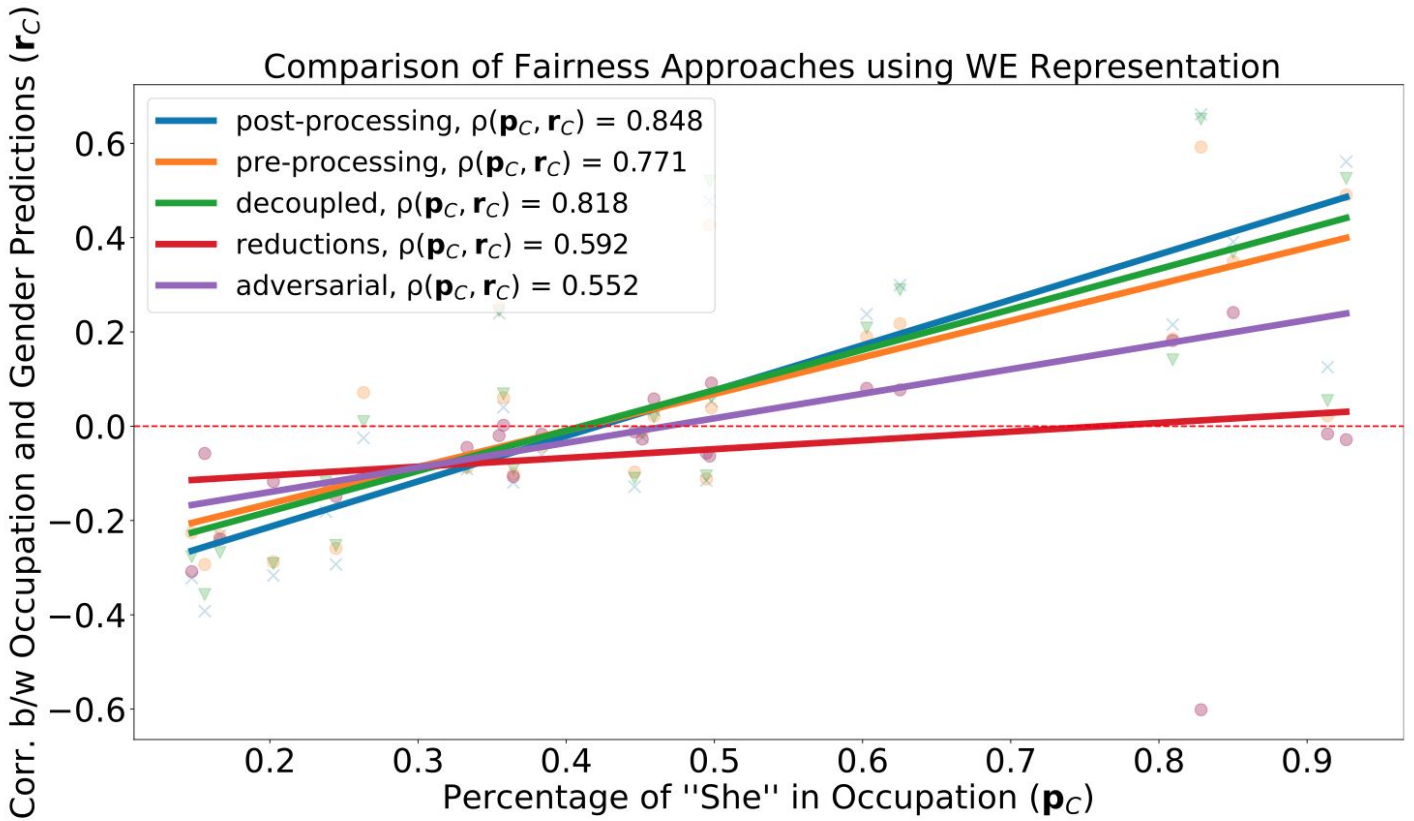

SNoB using WE Representation
Corr. b/w Occupation and Gender Predictions ($\mathbf{r}_c$)
Percentage of ''She'' in Occupation ($\mathbf{p}_c$)
0.6
0.4
0.2
0.0
−0.2
−0.4
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
dietitian
model
physician
yoga_teacher
psychologist
dentist
teacher
interior_designer
nurse
photographer
chiropractor
painter
attorney
poet
filmmaker
professor
architect
accountant
journalist
composer
pastor
personal_trainer
surgeon
comedian
software_engineer
post-processing, ρ($\mathbf{p}_c$, $\mathbf{r}_c$) = 0.848



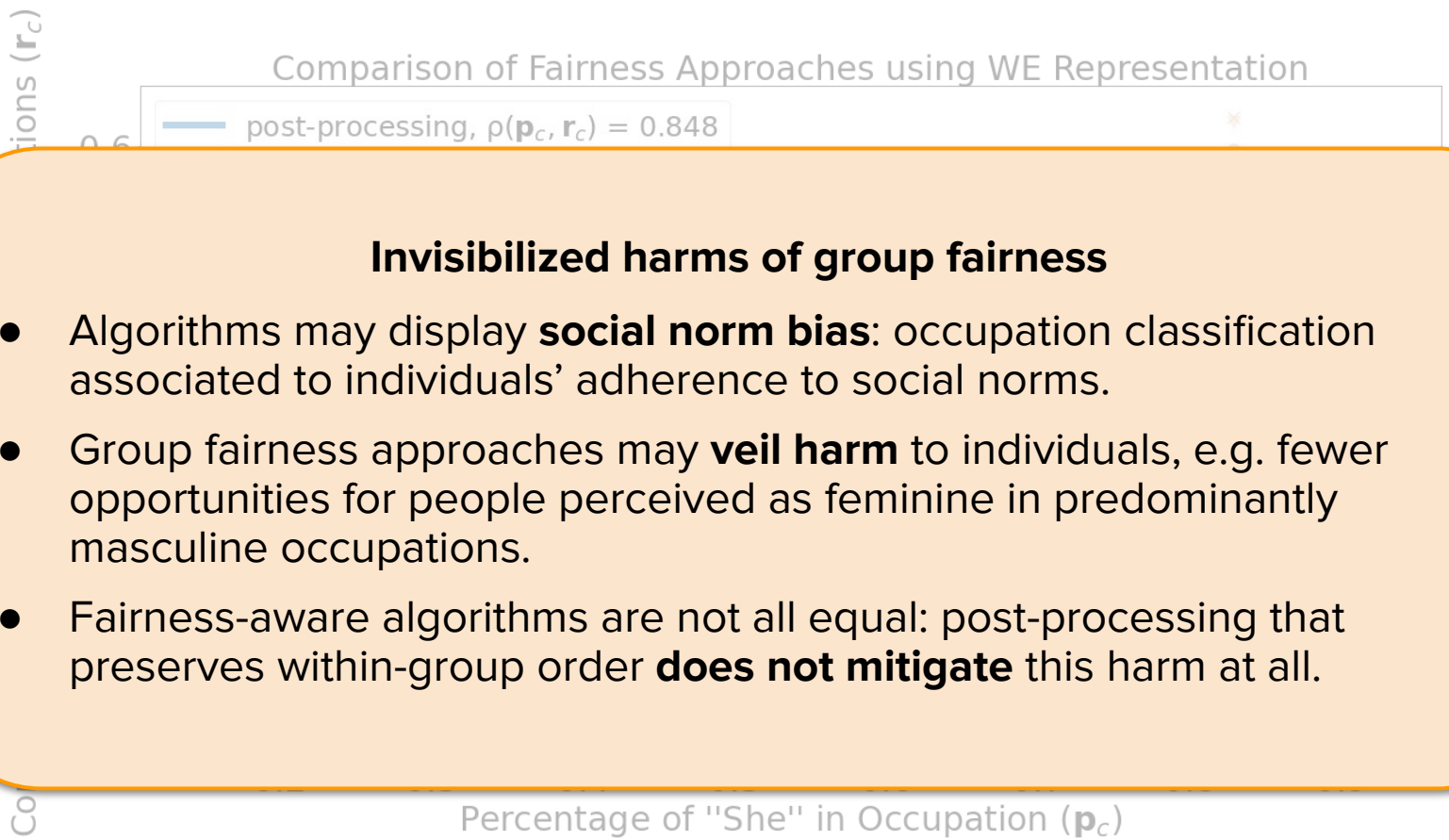
women surgeons and software engineers are less likely to be predicted as such if their bio "sound like a woman's"



Strength of correlation increases with occupation's gender imbalance

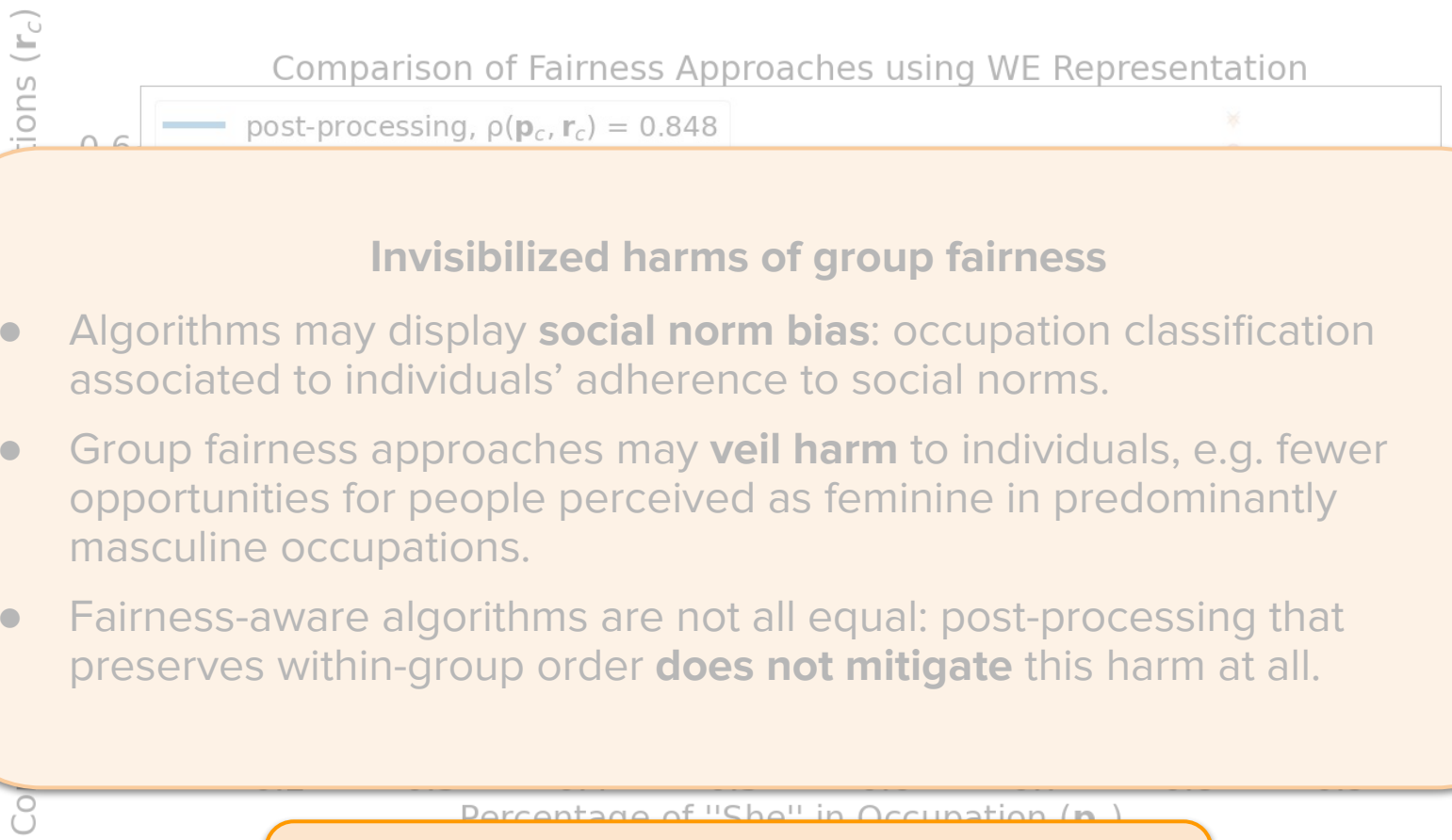


Gender correlations persist using “fair” interventions



Invisibilized harms of group fairness

- Algorithms may display **social norm bias**: occupation classification associated to individuals' adherence to social norms.
- Group fairness approaches may **veil harm** to individuals, e.g. fewer opportunities for people perceived as feminine in predominantly masculine occupations.
- Fairness-aware algorithms are not all equal: post-processing that preserves within-group order **does not mitigate** this harm at all.



Invisibilized harms of group fairness

- Algorithms may display **social norm bias**: occupation classification associated to individuals' adherence to social norms.
- Group fairness approaches may **veil harm** to individuals, e.g. fewer opportunities for people perceived as feminine in predominantly masculine occupations.
- Fairness-aware algorithms are not all equal: post-processing that preserves within-group order **does not mitigate** this harm at all.

Thank you!