PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

May 7, 2024

Rush T. Stewart (2022). *Identity and the limits of fair assessment*. Journal of Theoretical Politics 2022, Vol. 34(3), pp. 415 - 442.

Setup

A single property y of interest.

Individuals in N either have property y or lack it: $Y : N \to \{0, 1\}$

Call a function $h : N \to [0, 1]$ an assessor.

For concreteness, interpret $h(i)$ as the assessor's probability that i has property y.

Setup

The quantity $P(Y = 1) = \mu$, for example, is the proportion of people in N that have property y, the prevalence of y in the population.

Call μ the base rate for y in N.

Given a **partition** $\pi = \{G_1, \ldots, G_m\}$ of N, let $P_k = P(\cdot \mid G_k)$. So, $P_1(Y = 1) = \mu_1$ is the base rate for y in group 1 is μ_1 and $P_2(h = 0.5)$ is the proportion of people to which h assigns 0.5 in G_2, and so on.

Calibration/Predictive Equity

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

Calibration/Predictive Equity

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

An assessor h satisfies **predictive equity** (also called **weak calibration for groups**) for a partition π if

$$P_k(Y = 1 \mid h = p) = P_j(Y = 1 \mid h = p) \text{ for all } G_k, G_j \in \pi$$

Calibration Across 2 Groups

Let $N = \{1, 2, 3, 4, 5, 6\}$, and let $Y(i) = 1$ for $i = 1, 5, 6$. Consider the following two partitions

1. $\{B, W\} = \{\{1, 2, 4\}, \{3, 5, 6\}\}$ and
2. $\{M, F\} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$

	B	W
M	$h(1^*) = \frac{1}{2}, h(2) = 0$	$h(3) = \frac{1}{2}$
F	$h(4) = \frac{1}{2}$	$h(5^*) = \frac{1}{2}, h(6^*) = 1$

Calibration Across 2 Groups

Let $N = \{1, 2, 3, 4, 5, 6\}$, and let $Y(i) = 1$ for $i = 1, 5, 6$. Consider the following two partitions

1. $\{B, W\} = \{\{1, 2, 4\}, \{3, 5, 6\}\}$ and
2. $\{M, F\} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$

	B	W
M	$h(1^*) = \frac{1}{2}, h(2) = 0$	$h(3) = \frac{1}{2}$
F	$h(4) = \frac{1}{2}$	$h(5^*) = \frac{1}{2}, h(6^*) = 1$

h is calibrated across $\{M, F\}$ *and* across $\{B, W\}$.

No Calibration Across Groups

Suppose $N = \{1, 2, 3\}$ with $Y(1) = Y(3) = 1$ and $Y(2) = 0$:

	B	W
M	1^*	2
F	3^*	

No Calibration Across Groups

Suppose $N = \{1, 2, 3\}$ with $Y(1) = Y(3) = 1$ and $Y(2) = 0$:

	B	W
M	1^*	2
F	3^*	

- Supposing that h is imperfect and calibrated for the $\{M, F\}$ partition of N implies that individual 1 must receive a score in $(0, 1)$.
- The only such assessment consistent with calibration is $h(1) = h(2) = 1/2$.
- But then h cannot calibrated for B since, by calibration for F, $h(3) = 1$.
- Similarly, h cannot be calibrated for W since $P_W(Y = 1 \mid h = 1/2) \neq 1/2$.

- One one hand, requiring the satisfaction of a fairness constraint for some single partition is generally unsatisfactory since we may care about the fair treatment of groups from different partitions.

- One one hand, requiring the satisfaction of a fairness constraint for some single partition is generally unsatisfactory since we may care about the fair treatment of groups from different partitions.

- On the other hand, requiring any of the fairness constraints considered here be satisfied for *all* partitions of the population or all partitions of some cardinality places unrealistically high demands on assessment.

We cannot insist on any notion of statistical fairness for every subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. This simply corresponds to 'overfitting' a fairness constraint.

Michael Kearns, Seth Neel, Aaron Roth, Zhiwei, and Steven Wu (2018). *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*. In: Proceedings of the 35th International Conference on Machine Learning, Volume 80, Stockholm, Sweden, pp. 2564 - 2572. PMLR.

Intersectionality

Kimberlé Crenshaw, who introduced the term "intersectionality," makes use of a court case to explain how bias against black women, for example, is consistent with the lack of that form of bias against black people or against women.

Intersectionality

Kimberlé Crenshaw, who introduced the term "intersectionality," makes use of a court case to explain how bias against black women, for example, is consistent with the lack of that form of bias against black people or against women.

In *DeGraffenreid v. General Motors*, five black women alleging discrimination by General Motor's seniority-based system sued the company. Prior to 1964, General Motors did not hire black women. All of the black women hired after 1970 lost their jobs through a seniority-based layoff during a later recession.

K. Crenshaw (1989). *Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics*. University of Chicago Legal Forum 1989(Article 8): 139-167.

Intersectionality

The district court rejected the plaintiffs' attempt to bring a suit on behalf of black women in particular rather than on behalf of black people or women. According to the court, the suit must present "a cause of action for race discrimination, sex discrimination, or alternatively either, but not a combination of both".

Intersectionality

The district court rejected the plaintiffs' attempt to bring a suit on behalf of black women in particular rather than on behalf of black people or women. According to the court, the suit must present "a cause of action for race discrimination, sex discrimination, or alternatively either, but not a combination of both".

The court noted that, while General Motors did not hire black women prior to 1964, they did hire female employees for a number of years prior to 1964. So there was no sex discrimination.

Intersectionality

And what if General Motors had hired black people—specifically black men—for a number of years prior to 1964? Crenshaw's point is that that would not really absolve General Motors of the charge of discrimination against black women. It certainly does not follow that there could be no discrimination against black women.

Intersectional Bias

Let $N = \{1, 2, 3, 4, 5, 6, 7\}$, and let $Y(i) = 1$ for $i = 2, 4, 5, 7$. Consider two binary partitions $\{M, F\}$ and $\{B, W\}$.

	B	W
M	$h(1) = h(2^*) = \frac{2}{3}$	$h(3) = 0, h(4^*) = \frac{2}{3}$
F	$h(5^*) = \frac{2}{3}$	$h(6) = h(7^*) = \frac{2}{3}$

- The assessor h is calibrated for both the $\{M, F\}$ partition and the $\{B, W\}$ partition. In all of those groups, two thirds of those who receive an assessment of $2/3$ have property y.

Intersectional Bias

	B	W
M	$h(1) = h(2^*) = \frac{2}{3}$	$h(3) = 0, h(4^*) = \frac{2}{3}$
F	$h(5^*) = \frac{2}{3}$	$h(6) = h(7^*) = \frac{2}{3}$

- The coarsest common refinment is the four-cell partition $\{B \cap M, B \cap F, W \cap M, W \cap F\}$ composed of the groups of black men, black women, white men, and white women.

Intersectional Bias

	B	W
M	$h(1) = h(2^*) = \frac{2}{3}$	$h(3) = 0, h(4^*) = \frac{2}{3}$
F	$h(5^*) = \frac{2}{3}$	$h(6) = h(7^*) = \frac{2}{3}$

- The coarsest common refinment is the four-cell partition $\{B \cap M, B \cap F, W \cap M, W \cap F\}$ composed of the groups of black men, black women, white men, and white women.
- Since $P_{B \cap F}(Y = 1) \mid h = 2/3) = 1$, h is underconfident in (and so not calibrated for) black women. At the same time, h is overconfident in both black men and white women.

Observation 6. Let h be an assessor for N.

1. Even if h is calibrated for each partition in a set Π of partitions of N, h can fail to be calibrated for the *coarsest common refinement* of Π.
2. Even if h satisfies predictive equity for each partition in a set Π of partitions of N, h can fail to satisfy predictive equity for the *coarsest common refinement* of Π.

Observation 6. Let h be an assessor for N.

1. Even if h is calibrated for each partition in a set Π of partitions of N, h can fail to be calibrated for the *coarsest common refinement* of Π.
2. Even if h satisfies predictive equity for each partition in a set Π of partitions of N, h can fail to satisfy predictive equity for the *coarsest common refinement* of Π.

Observation 7. Let h be an assessor for N.

1. If h is calibrated for the *coarsest common refinement* of a set Π of partitions of N, then h is calibrated for each partition in Π.
2. If h satisfies predictive equity for the *coarsest common refinement* of a set Π of partitions of N, then h satisfies predictive equity for each partition in Π.

Concluding Remarks

There are multiple ways to carve a population, multiple social identities, for which it may be important to avoid biased assessments. Fixing a single partition of identities is overly restrictive, committing us to ignoring both relevant forms of bias against other groups and changing social context. Allowing even a set of partitions to ossify into *the* relevant partitions may fail to make us sufficiently attentive.

Concluding Remarks

There are multiple ways to carve a population, multiple social identities, for which it may be important to avoid biased assessments. Fixing a single partition of identities is overly restrictive, committing us to ignoring both relevant forms of bias against other groups and changing social context. Allowing even a set of partitions to ossify into *the* relevant partitions may fail to make us sufficiently attentive.

Where does this leave us? What the foregoing analysis helps us to make clear is that, not only is there a conflict between eliminating different forms of bias, but there are serious limits to the extent to which a given form of bias can be eliminated across different partitions.

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A clarification of the nuances in the fairness metrics landscape

Alessandro Castelnovo[1,2,3], Riccardo Crupi[1,3], Greta Greco[1,2,3], Daniele Regoli[1,3]✉, Ilaria Giuseppina Penco[1] & Andrea Claudio Cosentini[1]

In recent years, the problem of addressing fairness in machine learning (ML) and automatic decision making has attracted a lot of attention in the scientific communities dealing with artificial intelligence. A plethora of different definitions of fairness in ML have been proposed, that consider different notions of what is a "fair decision" in situations impacting individuals in the population. The precise differences, implications and "orthogonality" between these notions have not yet been fully analyzed in the literature. In this work, we try to make some order out of this zoo of definitions.

https://www.nature.com/articles/s41598-022-07939-1

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl (2024). *Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making*. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24).

Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III (2024). *The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features*. In 29th International Conference on Intelligent User Interfaces (IUI '24).

"[C]ompanies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums."

Jennifer Kite-Powell (2022). *Explainable AI is trending and here's why*. Forbes.

"[C]ompanies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums."

Jennifer Kite-Powell (2022). *Explainable AI is trending and here's why*. Forbes.

Explanations "...provide a more effective interface for the human-in-the-loop, enabling people to identify and address fairness and other issues"

Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan (2019). *Explaining models: An empirical study of how explanations impact fairness judgment*. In Proceedings of the 24th International Conference on Intelligent User Interfaces. 275–285.

In order for a human-in-the-loop to addresses fairness issues, they should have the capacity to identify mistaken recommendations, reducing the false negative errors affecting that group.

In this case, the goal of explanations should be to help humans identify such errors, yielding AI-assisted decisions that have better distributive fairness properties than the AI alone.

Note that this is different to the perceptions that humans may have of an AI system, and it also differs from the overall accuracy or reliance behavior.

Fairness through Unawareness

"Fairness through unawareness": an AI system is fair if it does not make use of information that is evidently indicative of a person's demographics.

- Neither a sufficient nor a necessary condition for fairness

Sam Corbett-Davies and Sharad Goel (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv preprint arXiv:1808.00023.

Explanation

Focus on *feature-based explanations*: LIME is used in the experiments, due to its popularity in the literature as well as in practice and, importantly, the fact that LIME has been claimed to enable fairness assessments.

Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli (2020). *LimeOut: An ensemble approach to improve process fairness*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 475 - 491.

Joymallya Chakraborty, Kewen Peng, and Tim Menzies (2020). *Making fair ML software using trustworthy explanation*. In Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, pp. 1229-1233.

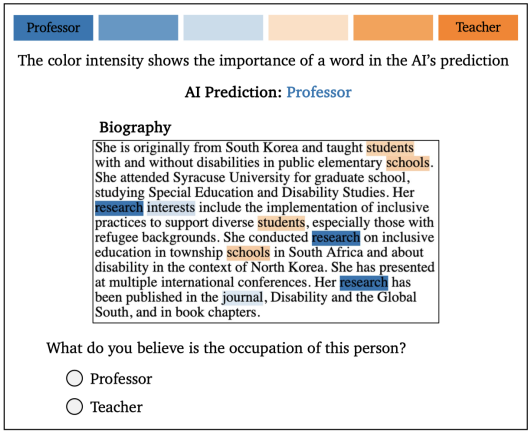
Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online. An important task herein is to determine someone's occupation, which is a prerequisite for advertising job openings or recruiting people for adequate positions. This information may not be readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI systems, it is susceptible to gender bias and discrimination.

Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online. An important task herein is to determine someone's occupation, which is a prerequisite for advertising job openings or recruiting people for adequate positions. This information may not be readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI systems, it is susceptible to gender bias and discrimination.

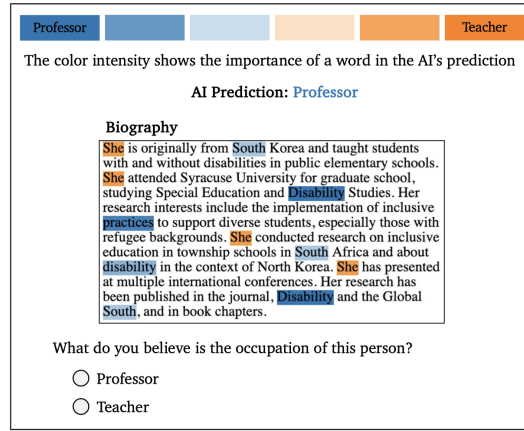
This study: instantiate an AI-assisted decision-making setup where participants see short textual bios and are asked—with the help of an AI recommendation to predict whether a given bio belongs to a professor or a teacher. Professors are historically a men-dominated occupation, whereas teachers have been mostly associated with women.

Experiment

- Each participant sees 14 bios one by one, each including the AI recommendation as well as an explanation highlighting the most predictive words. We also include a baseline condition without explanations.
- Participants are assigned to conditions where they see recommendations and explanations either from (i) an AI model that uses task-relevant features, or (ii) an AI model that uses gendered (i.e., sensitive) features.
- Participants in each condition first complete the task of predicting occupations for 14 bios, and—if assigned to a condition with explanations—answer several questions regarding their fairness perceptions.



***Task-relevant* condition**



***Gendered* condition**