Fairness, equality, and power in algorithmic decision making

Rediet Abebe Maximilian Kasy

February 2021

Introduction

- Public debate and the computer science literature:
  Fairness of algorithms, understood as the absence of **discrimination**.
- We argue: Leading definitions of fairness have three limitations:
  1. They legitimize inequalities justified by "merit."
  2. They are narrowly bracketed; only consider differences of treatment within the algorithm.
  3. They only consider between-group differences.
- Two alternative perspectives:
  1. What is the causal impact of the introduction of an algorithm on **inequality**?
  2. Who has the **power** to pick the objective function of an algorithm?

Fairness in algorithmic decision making – Setup

- Binary treatment W, treatment return M (heterogeneous), treatment cost c.
  Decision maker's objective

$$\mu = E[W \cdot (M - c)].$$

- All expectations denote averages across individuals (not uncertainty).

- M is unobserved, but predictable based on features X.
  For $m(x) = E[M|X = x]$, the optimal policy is

$$w^*(x) = \mathbf{1}(m(X) > c).$$

Examples

- Bail setting for defendants based on predicted recidivism.
- Screening of job candidates based on predicted performance.
- Consumer credit based on predicted repayment.
- Screening of tenants for housing based on predicted payment risk.
- Admission to schools based on standardized tests.

Fairness

Inequality

Power

Case study

Definitions of fairness

- Most definitions depend on **three ingredients**.
  1. Treatment W (job, credit, incarceration, school admission).
  2. A notion of merit M (marginal product, credit default, recidivism, test performance).
  3. Protected categories A (ethnicity, gender).

- I will focus, for specificity, on the following **definition of fairness**:

$$\pi = E[M|W = 1, A = 1] - E[M|W = 1, A = 0] = 0$$

"Average merit, among the treated, does not vary across the groups a."

This is called "predictive parity" in machine learning,
the "hit rate test" for "taste based discrimination" in economics.

- "Fairness in machine learning" literature: **Constrained optimization**.

$$w^*(\cdot) = \underset{w(\cdot)}{\operatorname{argmax}}\, E[w(X) \cdot (m(X) - c)] \qquad \text{subject to} \qquad \pi = 0.$$

Fairness and $\mathscr{D}$'s objective

Observation

Suppose that W, M are binary ("classification"), and that

1. $m(X) = M$ *(perfect predictability), and*
2. $w^*(x) = \mathbf{1}(m(X) > c)$ *(unconstrained maximization of $\mathscr{D}$'s objective μ).*

Then $w^*(x)$ satisfies predictive parity, i.e., $\pi = 0$.

In words:

- If $\mathscr{D}$ is a firm that is maximizing profits and observes everything then their decisions are fair by assumption.
  - No matter how unequal the resulting outcomes within and across groups.
- Only deviations from profit-maximization are "unfair."

Three normative limitations of “fairness” as predictive parity

1. They legitimize and perpetuate **inequalities justified by “merit.”**
   Where does inequality in M come from?

2. They are **narrowly bracketed**.
   Inequality in W in the algorithm,
   instead of some outcomes Y in a wider population.

3. Fairness-based perspectives **focus on categories** (protected groups)
   and ignore within-group inequality.

$\Rightarrow$ We consider the impact on inequality or welfare as an alternative.

Three normative limitations of “fairness” as predictive parity

1. They legitimize and perpetuate **inequalities justified by “merit.”**
   Where does inequality in M come from?

2. They are **narrowly bracketed**.
   Inequality in W in the algorithm,
   instead of some outcomes Y in a wider population.

3. Fairness-based perspectives **focus on categories** (protected groups) and ignore within-group inequality.

$\Rightarrow$ We consider the impact on inequality or welfare as an alternative.

Three normative limitations of “fairness” as predictive parity

1. They legitimize and perpetuate **inequalities justified by “merit.”**
   Where does inequality in M come from?

2. They are **narrowly bracketed**.
   Inequality in W in the algorithm,
   instead of some outcomes Y in a wider population.

3. Fairness-based perspectives **focus on categories** (protected groups)
   and ignore within-group inequality.

$\Rightarrow$ We consider the impact on inequality or welfare as an alternative.

Fairness

Inequality

Power

Case study

The impact on inequality or welfare as an alternative

- Outcomes are determined by the **potential outcome equation**

$$Y = W \cdot Y^1 + (1 - W) \cdot Y^0.$$

- The **realized outcome** distribution is given by

$$p_{Y,X}(y,x) = \left[p_{Y^0|X}(y,x) + w(x) \cdot \left(p_{Y^1|X}(y,x) - p_{Y^0|X}(y,x) \right) \right] \cdot p_X(x).$$

- What is the impact of $w(\cdot)$ on a **statistic** ν?

$$\nu = \nu(p_{Y,X}).$$

Examples: Variance, quantiles, between group inequality.

When fairness and equality are in conflict

- Fairness is about **treating** people of the same "**merit**" independently of their **group** membership.
- Equality is about the (counterfactual / causal) **consequences** of an algorithm for the distribution of **welfare** of different **people**.

Examples when they are in conflict:

1. Increased surveillance / **better prediction** algorithms:
   Lead to treatments more aligned with "merit"
   God for fairness, bad for equality.
2. Affirmative action / **compensatory interventions** for pre-existing inequalities:
   Bad for fairness, good for equality.

Power

- Both fairness and equality are about differences between people who are **being treated**.
- Elephant in the room:
  - Who is on the **other side** of the algorithm?
  - Who gets to be the decision maker $\mathscr{D}$ – who gets to pick the objective function μ?
- Political economy perspective:
  - **Ownership of the means of prediction**.
  - Data and algorithms.

Fairness

Inequality

Power

Case study

Case study

- Compas risk score data for recidivism.
- From Pro-Publica's reporting on algorithmic discrimination in sentencing.

Mapping our setup to these data:

- A: race (Black or White),
- W: risk score exceeding 4,
- M: recidivism within two years,
- Y: jail time,
- X: race, sex, age, juvenile counts of misdemeanors, fellonies, and other infractions, general prior counts, as well as charge degree.

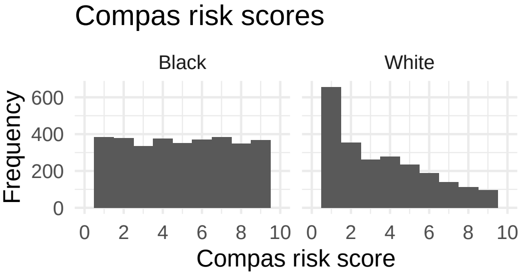
Counterfactual scenarios

Compare three scenarios:

1. "Affirmative action:" Adjust risk scores ± 1, depending on race.
2. Status quo.
3. Perfect predictability: Scores equal 10 or 1, depending on recidivism in 2 years.

For each: Impute counterfactual

- W: Counterfactual score bigger than 4.
- Y: Based on a causal-forest estimate of the impact on Y of risk scores, conditional on the covariates in X.
- This relies on the assumption of conditional exogeneity of risk-scores given X. Not credible, but useful for illustration.


Compas risk scores
Black
White
Frequency
600
400
200
0
0
2
4
6
8
10
Compas risk score

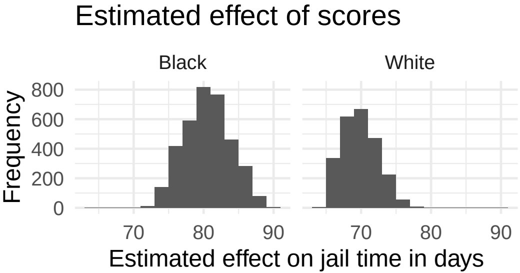

Estimated effect of scores
Black
White
Frequency
800
600
400
200
0
70
80
90
Estimated effect on jail time in days

Table: Counterfactual scenarios, by group

	Black			White		
Scenario	(Score>4)	Recid|(Score>4)	Jail time	(Score>4)	Recid|(Score>4)	Jail time
Aff. Action	0.49	0.67	49.12	0.47	0.55	36.90
Status quo	0.59	0.64	52.97	0.35	0.60	29.47
Perfect predict.	0.52	1.00	65.86	0.40	1.00	42.85

Table: Counterfactual scenarios, outcomes for all

Scenario	Score>4	Jail time	IQR jail time	SD log jail time
Aff. Action	0.48	44.23	23.8	1.81
Status quo	0.49	43.56	25.0	1.89
Perfect predict.	0.48	56.65	59.9	2.10

Thank you!