Fairness, equality, and power in algorithmic decision making

Rediet Abebe Maximilian Kasy

May 2020

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

- 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.

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."

- "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

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., $\pi = 0$.

Reasons for bias

1. **Preference-based** discrimination.
   The decision maker is maximizing some objective other than μ.

2. **Mis-measurement** and biased beliefs.
   Due to bias of past data, $m(X) \neq E[M|X]$.

3. **Statistical discrimination**.
   Even if $w^*(\cdot) = \text{argmax } \pi$ and $m(X) = E[M|X]$,
   $w^*(\cdot)$ might violate fairness if X does not perfectly predict M.

Three limitations of “fairness” perspectives

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 limitations of “fairness” perspectives

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 limitations of "fairness" perspectives

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

Examples

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

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

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

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

- Examples:
  - Variance $\mathrm{Var}(Y)$,
  - "welfare" $E[Y^\gamma]$,
  - between-group inequality $E[Y|A = 1] - E[Y|A = 0]$.

Influence function approximation of the statistic ν

$$\nu(p_{Y,X}) - \nu(p^*_{Y,X}) \approx E[IF(Y,X)],$$

- $IF(Y,X)$ is the influence function of $\nu(p_{Y,X})$.
- The expectation averages over the distribution $p_{Y,X}$.
- Examples:

$$\begin{aligned}
\nu &= E[Y] & IF &= Y - E[Y] \\
\nu &= \mathrm{Var}(Y) & IF &= (Y - E[Y])^2 - \mathrm{Var}(Y) \\
\nu &= E[Y|A=1] - E[Y|A=0] & IF &= Y \cdot \left(\frac{A}{E[A]} - \frac{1-A}{1-E[A]} \right).
\end{aligned}$$

The impact of marginal policy changes on profits, fairness, and inequality

Proposition

Consider a family of assignment policies $w(x) = w^*(x) + \epsilon \cdot dw(x)$. *Then*

$$d\mu = E[dw(X) \cdot l(X)], \qquad d\pi = E\left[dw(X) \cdot p(X)\right], \qquad d\nu = E[dw(X) \cdot n(X)],$$

where

$$l(X) = E[M|X = x] - c, \tag{1}$$

$$p(X) = E\left[(M - E[M|W = 1, A = 1]) \cdot \frac{A}{E[WA]} \right.$$

$$\left. - \qquad (M - E[M|W = 1, A = 0]) \cdot \frac{(1 - A)}{E[W(1 - A)]} \middle| X = x \right], \tag{2}$$

$$n(x) = E\left[IF(Y^1, x) - IF(Y^0, x)|X = x\right]. \tag{3}$$

Power

- Recap:
  1. Fairness: Critique the unequal treatment of **individuals** i who are of the same merit M. Merit is defined in terms of $\mathscr{D}$'s objective.
  2. Equality: Causal impact of an algorithm on the distribution of relevant outcomes Y across **individuals** i more generally.
- 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.

Implied welfare weights

- What welfare weights would rationalize actually chosen policies as optimal?
- That is, in who's interest are decisions being made?

Corollary

Suppose that welfare weights are a function of the observable features X, and that there is again a cost of treatment c. A given assignment rule $w(\cdot)$ is a solution to the problem

$$\underset{w(\cdot)}{\operatorname{argmax}}\ E[w(X) \cdot (\omega(X) \cdot E[Y^1 - Y^0|X] - c)]$$

if and only if

$$\begin{aligned} w(x) = 1 &\Rightarrow \omega(X) > c/E[Y^1 - Y^0|X]) \\ w(x) = 0 &\Rightarrow \omega(X) < c/E[Y^1 - Y^0|X]) \\ w(x) \in]0,1[&\Rightarrow \omega(X) = c/E[Y^1 - Y^0|X]). \end{aligned}$$

Fairness

Inequality

Power

Examples

Case study

Example of limitation 1: Improvement in the predictability of merit.

- Limitation 1: Fairness legitimizes inequalities justified by "merit."
- Assumptions:
  - Scenario a: The decisionmaker only observes A.
  - Scenario b: They can perfectly predict (observe) M based on X.
  - $Y = W$, M is binary with $P(M = 1|A = a) = p^a$, where $0 < c < p^1 < p^0$.
- Under these assumptions

$$W^a = \mathbf{1}(E[M|A] > c) = 1, \qquad W^b = \mathbf{1}(E[M|X] > c) = M.$$

- Consequences:
  - The policy a is unfair, the policy b is fair. $\pi_a = p^1 - p^0$, $\pi_b = 0$.
  - Inequality of outcomes has increased.

$$\mathrm{Var}_a(Y) = 0, \qquad \mathrm{Var}_b(Y) = E[M](1 - E[M]) > 0.$$

  - Expected welfare $E[Y^\gamma]$ has decreased.

$$E_a[Y^\gamma] = 1, \qquad E_b[Y^\gamma] = E[M] < 1.$$

Example of limitation 2: A reform that abolishes affirmative action.

- Limitation 2: Narrow bracketing. Inequality in treatment W, instead of outcomes Y.
- Assumptions:
  - Scenario a: The decisionmaker receives a subsidy of 1 for hiring members of the group $A = 1$.
  - Scenario b: They subsidy is abolished
  - (M, A) is uniformly distributed on $\{0, 1\}^2$, M is perfectly observable, $0 < c < 1$.
  - Potential outcomes are given by $Y^w = (1 - A) + w$.
- Under these assumptions

$$W^a = \mathbf{1}(M + A \geq 1), \qquad W^b = M.$$

- Consequences:
  - The policy a is unfair, the policy b is fair. $\pi_a = -.5$, $\pi_b = 0$.
  - Inequality of outcomes has increased.

$$\mathrm{Var}_a(Y) = 3/16, \qquad \mathrm{Var}_b(Y) = 1/2,$$

  - Expected welfare $E[Y^\gamma]$ has decreased.

$$E_a[Y^\gamma] = .75 + .25 \cdot 2^\gamma, \qquad E_b[Y^\gamma] = .5 + .25 \cdot 2^\gamma.$$

Example of limitation 3: A reform that mandates fairness.

- Limitation 3: Fairness ignores within-group inequality.
- Assumptions:
  - Scenario a: The decisionmaker is unconstrained.
  - Scenario b: They decisionmaker has to maintain fairness, $\pi = 0$.
  - $P(A = 1) = .5$, $c = .7$,

$$M|A = 1 \sim Unif(\{0, 1, 2, 3\}) \qquad M|A = 0 \sim Unif(\{1, 2\}).$$

  - Potential outcomes are given by $Y^w = M + w$.
- Under these assumptions

$$W^a = \mathbf{1}(M \geq 1), \qquad W^b = \mathbf{1}(M + A \geq 2).$$

- Consequences:
  - The policy a is unfair, the policy b is fair. $\pi_a = .5$, $\pi_b = 0$.
  - Inequality of outcomes has increased.

$$\mathrm{Var}_a(Y) = 1.234375, \qquad \mathrm{Var}_b(Y) = 2.359375,$$

  - Expected welfare $E[Y^\gamma]$ has decreased. For $\gamma = .5$,

$$E_a[Y^\gamma] = 1.43, \qquad E_b[Y^\gamma] = 1.08.$$

Fairness

Inequality

Power

Examples

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.

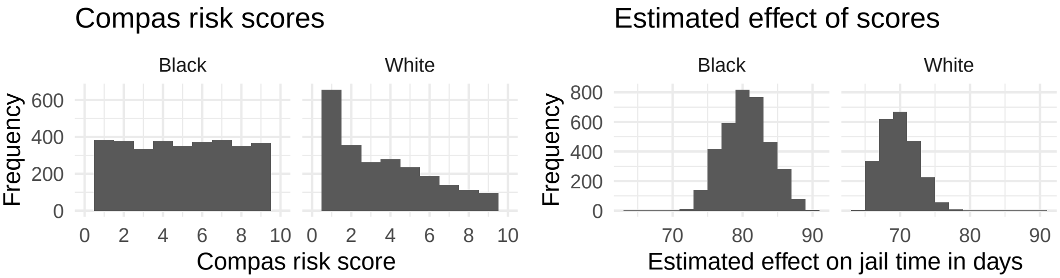
Compas risk scores
Black
White
Frequency
600
400
200
0
0 2 4 6 8 10
Compas risk score
Estimated effect of scores
Black
White
Frequency
800
600
400
200
0
70 80 90
Estimated effect on jail time in days

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!