Fairness, equality, and power in algorithmic decision making

Maximilian Kasy

March 2020

Areas of research that I am currently working on

- Theory of adaptive experimental design. (Department seminar on Thursday.)
  - Effect estimation, participant welfare, policy choice, or utilitarian welfare.
  - Related to active learning in AI.
- Actual field experiments.
  - Job search assistance for refugees in Amman & Irdib, Jordan.
  - Job guarantee pilot in Mariental, Austria.
  - Basic income in Marica, Brazil.
- Statistics in a social context.
  - Publication bias and optimal publication rules.
  - A theory of pre-analysis plans as commitment devices.
- Statistical theory of supervised machine learning.
  - Cross-validation, approximate cross-validation, analytical risk estimators.
- Ethics, justice and political economy of AI.
  - **This talk** – work in progress joint with Rediet Abebe.
  - Motivated by limitations of current debates about fairness in AI.

Fairness in algorithmic decision making

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

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

- 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 based on predicted recidivism.
  - Consumer credit based on predicted repayment.
  - 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**.

$$w^*(\cdot) = \underset{w(\cdot)}{\operatorname{argmax}}\ \mu = E[w(X) \cdot (m(X) - c)] \qquad \text{subject to}$$
$$\pi = E[M|W = 1, A = 1] - E[M|W = 1, A = 0] = 0.$$

Sources and limitations of (un)fairness

- Three 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 $m(X) = E[M|X]$,
     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 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.

The impact on inequality or welfare as an alternative

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

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

- **Realized outcome** distribution

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

- 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 to 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

$$\begin{aligned}
l(X) &= E[M|X = x] - c, && (1)\\
p(X) &= E\Bigg[(M - E[M|W = 1, A = 1]) \cdot \frac{A}{E[WA]} \\
&- \qquad (M - E[M|W = 1, A = 0]) \cdot \frac{(1 - A)}{E[W(1 - A)]}\Bigg|X = x\Bigg], && (2)\\
n(x) &= E\left[IF(Y^1, x) - IF(Y^0, x)|X = x\right]. && (3)
\end{aligned}$$

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: Narrowly 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 $\gamma = .5$,

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

Outlook

- Further characterizations when fairness and equality do / do not have the same implications.
- Empirical applications. Suggestions?
- Elaborating a third alternative perspective: Power.
  - Who gets to pick the objective function π?
  - Is maximization of ad-clicks really the socially most beneficial use of AI?
  - For given algorithmic decisions, what are the implied welfare weights that would rationalize these algorithms?

Thank you!