Fairness, equality, and power in algorithmic decision making

Maximilian Kasy

January 3, 2020

In the news.

There's software used across the country to predict future criminals. And it's biased against blacks.

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

Paperclip-making robots 'wipe out humanity' in killer AI Doomsday experiment

Introduction

- Algorithmic decision making in consequential settings:
  Hiring, consumer credit, bail setting, news feed selection, pricing, ...
- Public concerns:

  Are algorithms discriminating?
  Can algorithmic decisions be explained?
  Does AI create unemployment?
  What about privacy?

- Taken up in computer science:

  "Fairness, Accountability, and Transparency,"
  "Value Alignment," etc.

- Normative foundations for these concerns?
  How to evaluate decision making systems empirically?
- Economists (among others) have debated related questions
  in non-automated settings for a long time!

Work in progress

- Kasy, M. and Abebe, R. (2020).
  Fairness, equality, and power in algorithmic decision making.
  Forthcoming, FAccT 2021
- Kasy, M. and Abebe, R. (2020).
  Multitasking, surrogate outcomes, and the alignment problem.
- Kasy, M. and Teytelboym, A. (2020).
  Adaptive combinatorial allocation.

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

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

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).

- We 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.

Observation

- 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.

Corresponding examples where assessments based on inequality conflict with fairness:

1. Increased surveillance or predictive capacity.
2. Affirmative action or compensatory interventions.
3. Non-discrimination mandates.

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, quantiles, between group inequality.

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*

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

where

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

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

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

where

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

Uses of the proposition

1. Elucidate the **tension** between objectives.
   - Profits vs. fairness vs. equality vs. welfare?
   - $\Rightarrow$ Characterizes which parts of the feature space drive the tension between alternative objectives.
2. Solve for **optimal assignment** subject to constraints.
   - E.g. maximize μ subject to $\pi = 0$.
   - Then $w(x) = \mathbf{1}(l(x) > \lambda p(x))$.
3. **Power and inverse welfare weights**
   - For a given $w(\cdot)$, what objective is implicitly maximized?
   - What are the weights for different individuals that rationalize $w(\cdot)$?
4. **Algorithmic auditing**.
   - Similar to distributional decompositions in labor economics.

Thank you!