Discussion of: “Towards a Non-Discriminatory Algorithm in Selected Data”

Maximilian Kasy

April 9, 2021

Summary

- Denote Y^* the latent binary "merit", R the defendant's race, $f(X)$ the risk score.
- Measure of bias:

$$\Delta = E[E[f(X)|R = w, Y^*] - E[f(X)|R = b, Y^*]]$$

- Identification problem:
  Y^* is only observed when $D = 1$ (defendant got released).
- Proposed solution:
  1. Instrument D, using random judge assignment.
  2. Extrapolate to the full population
     ("identification at infinity"; cf. Heckman selection correction).
  3. Show that identifying $E[Y^*|R]$ and $E[f(X) \cdot Y^*|R]$ is enough to get Δ.
- Additionally:
  Solve for the OLS predictor of Y^* given X
  subject to $\Delta = 0$.

Two follow-up questions

- I think this is a well executed, transparent analysis
  of selection bias and instrumentation,
- importing these ideas into debates about algorithmic bias,
  and ML more broadly.
- I will ask two follow-up questions:
  1. When does this identification problem arise in algorithmic decision making?

     Why has the ML literature not dealt with this?
  2. By what normative criteria should we evaluate automated decisionmaking systems?

     What questions should we ask about algorithms if our goal is to reduce
     racialized mass incarceration, poverty, educational disparities, etc.?

When the selection problem does / doesn't arise

- **Question**: Why has the machine learning literature engaged so little with questions of selection, instruments, causality?

- Even for targeted treatment assignment, multi-armed bandits, reinforcement learning, which are about the causal effect of actions?

- **Answer**: Focus on maximizing rewards.
  $\Rightarrow$ Algorithms care only about the causal effect of their own actions.

- Their own actions are by construction exogenous conditional on the information that they use.

- The selection problem only comes in when the actions of human actors, affect observability, based on unobserved information.

  $\Rightarrow$ Setting of the present paper!
  Important for any hybrid human-machine decision making!

Two approaches to questions of justice

1. **Just deserts** (e.g. Libertarianism):
   "Does everyone get what they deserve, based on their merit?"

2. **Consequentialism** (e.g. Utilitarianism, welfare economics):
   "How does this policy / algorithm impact the wellbeing of those affected?"

Becker, contra the civil rights movement,
defined taste based discrimination in the libertarian framework:

- Whatever competitive profit maximizing firms do is defined as just.
- No matter how much inequality or poverty results!
- Deviations from profit maximization are called taste based discrimination.
- Present work on discrimination / fairness in ML continues this tradition.
- Analogy here: Whatever judges maximizing incapacitation do is just.

Fairness versus equality

- 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 (cf. *Abebe and Kasy 2021*):

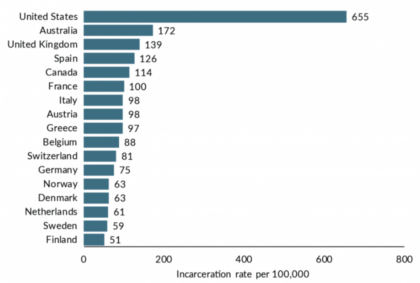
1. Increased surveillance / **better prediction** algorithms:
   Lead to treatments more aligned with "merit"
   Good for fairness, bad for equality.

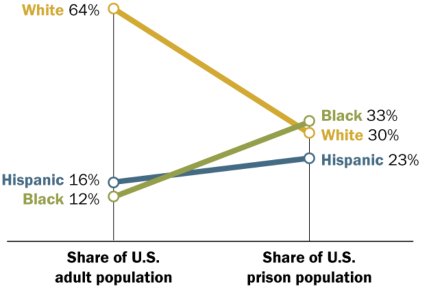
2. Affirmative action / **compensatory interventions** for pre-existing inequalities:
   Bad for fairness, good for equality.

A call to arms

What questions should we ask, as economists, if we want to help end racialized mass incarceration, poverty, educational disparities, etc.?

- Let's ask less about fairness:
  "Can we rationalize incarceration and its racial gaps based on different criminal propensity?"
- Let's ask more about consequences and inequality:
  "How does this algorithm / reform impact unequal incarceration, between and within racial groups?"





Thank you!