The social impact of algorithmic decision making: Economic perspectives

Maximilian Kasy

December 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.
- Kasy, M. and Abebe, R. (2020).
  Multitasking, surrogate outcomes, and the alignment problem.
- Kasy, M. and Teytelboym, A. (2020).
  Adaptive combinatorial allocation.

Introduction

Fairness, equality, and power in algorithmic decision making

- Fairness
- Inequality

Multi-tasking, surrogates, and the alignment problem

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.

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**: *"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.

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

Alternative approaches

- 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?
- **Tension** between objectives.
  - Profits vs. fairness vs. equality vs. welfare?
  - We characterize which parts of the feature space drive the tension between alternative objectives.
- We propose a standardized procedure for **algorithmic auditing**, estimating the causal impact of an algorithm on the distribution of welfare relevant outcomes.

Introduction

Fairness, equality, and power in algorithmic decision making

- Fairness
- Inequality

Multi-tasking, surrogates, and the alignment problem

The value alignment problem

- Much recent attention; e.g. Russell (2019):

  > *[...] we may suffer from a failure of value alignment—we may, perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own*

- The debate in tech & CS focuses on robotics and the grand, e.g. Bostrom (2003):

  > *Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. [...]*

Value alignment and observability

- No need to wait for the singularity:
  There are many **examples** (outside robotics) in the **social world**, right now.
  - Social media feeds maximizing clicks.
  - Teachers promoted based on student test scores.
  - Doctors paid per patient.
  - ...
- One unifying theme: Lack of **observability** of welfare.
- How to design
  - reward functions,
  - incentive systems,
  - adaptive treatment assignment algorithms,

  when our true objective is not observed?

Two related literatures to build on

1. **Multi-tasking in contract design**
   - Holmstrom and Milgrom (1991):
     Why are high-powered economic incentives rarely observed?
   - A: Because they would distort effort away from unobserved dimensions.
   - E.g., "Teaching to the test."
2. **Surrogate outcomes in clininical trials**
   - How to evaluate medical treatments, when relevant outcomes take too long to realize?
   - A: When we observe relevant variables on all causal pathways.
   - E.g., blood pressure and heart disease.

Thank you!