An Economic Approach to Regulating Algorithms

Ashesh Rambachan, Jon Kleinberg, Jens Ludwig & Sendhil Mullainathan

November 3, 2020
Motivation

- Algorithms are increasingly used in a wide variety of important domains.
- **Criminal justice**: Should a defendant be granted bail?
- **Medicine**: Should a patient be tested?
- **Hiring**: Should an applicant be hired?
- **Finance**: Should an applicant receive a loan?
- Widespread fears that algorithmic decision-making may reflect or worsen existing socioeconomic disparities.
Algorithmic Fairness: Perspective from Computer Science

- An algorithm produces a prediction function \( \hat{f} \), which predicts label \( Y^* \) from features \( W \).
  
  \[ \Rightarrow \] Define what it means for \( \hat{f} \) to be fair.
  
  \[ \Rightarrow \] Take chosen def’n of “fair prediction function” as primitive.

- Constructing fair algorithms reduces to introducing an additional constraint in our training procedures:

\[
\min_{f} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(Y^*_i, W_i; f) \quad \text{s.t.} \quad f \text{ is “fair.”}
\]

- Enormously influential literature: Dwork et al. (2012), Zemel et al. (2013), Hardt et al. (2016) and many, many others.
This Paper: An Economic Approach

- Embed concerns about algorithmic bias within a social welfare function.
  - Defined over outcomes generated by decisions.
  - Captures explicit preference for efficiency and equity.

- Explicit equity preference in SWF generates concern about possible algorithmic bias.

- Our Approach: begin w/ SWF and derive implications of equity preferences for algorithm construction.
  - Analogy to optimal tax: derive properties of tax system, taking SWF as primitive.
Our Approach: Cast questions surrounding design/use of algorithms and algorithmic fairness into canonical welfare economics framework.

Welfare economics framework highlights two distinct problem formulations.

First-Best Problem: Benevolent “social planner” has full control over design/use of algorithmic decision rule.
  - Statistical decision theory

Second-Best Problem: Third-party DMs control design/use of algorithmic decision rule, and do not share same objectives as society.
  - What are DMs maximizing? What information do they have? What information does the social planner have about DMs? What policy tools are available?
  - Contracting problem, Mechanism design
Our Approach: Cast questions surrounding design/use of algorithms and algorithmic fairness into canonical welfare economics framework.

Big Picture Goal: Provide framework for economists to think about algorithmic fairness, and thereby fruitfully collaborate w/ computer science community.

- Richer analyses of second-best/regulation problem requires tractable modelling of both statistical and economic aspects of the problem.
Screening decisions

**Screening Decisions**: Individuals screened into program based on prediction of unknown outcome of interest \( Y^* \in \{0, 1\} \)
- Common example of “prediction policy problem” (Kleinberg et al., 2015, 2018a)

**Examples** include pre-trial release, hiring decisions, credit approvals and more.

Population of individuals described by characteristics \( W \in \{0, 1\}^J \) and group membership \( G \in \{0, 1\} \).
- \( G = 1 \) for “protected group.”
- \( P(g, w) = \mathbb{P}\{G = g, W = w\} \), \( \theta^*(g, w) = \mathbb{E}[Y^* | G = g, W = w] \).

Predict \( Y^* \) given observed features \((G, W)\) and admit individuals based upon predictions.
What algorithm would social planner use?

- Consider social planner that constructs algorithm and selects admissions rule in the screening decision.
  - Equivalently: Benevolent private actor that shares society’s goals.

- This is the **first-best problem**. Analyze first-best to understand how equity preferences affect screening decisions.

- **Questions**: What algorithm would social planner construct? What admissions rule would social planner select?
The social welfare function

- Weighted average of expected outcome of interest among admitted individuals:

\[ \sum_{g,w} \psi_g \theta^*(g, w) t(g, w) P(g, w), \]

- Admissions rule \( t(g, w) \in [0, 1] \)
- Generalized social welfare weights, \( \psi_g \geq 0 \).

- SWF captures preference for both **efficiency** and **equity**
  - Efficiency: maximize expected outcome of interest among admitted group
  - Equity: value outcomes associated w/ protected group more, \( \psi_1 > \psi_0 \).

- Social planner does not know \( \theta^*(g, w) \) \( \implies \) faces non-trivial *prediction policy* problem.
Measured outcomes and the training dataset

- Social planner receives access to **training dataset** to construct predictions.
  - Training dataset $D_N$ consists of $N$ i.i.d. random draws from population.

- Each individual additionally described by **measured outcome** $\tilde{Y}$.
  - May differ from true outcome of interest $Y^*$.

- Training dataset useful provided measured outcome $\tilde{Y}$ informative about outcome of interest $Y^*$.
  - Specifies **prior beliefs** $\pi(\cdot)$ over $(Y^*, \tilde{Y})|G, W$.
  - Knows marginal dist $(G, W)$. 
The first-best screening problem

- Social planner faces known **capacity constraint** $C \in [0, 1]$. May not admit more than fraction $C$ of population into program.

- **First-best problem**: Maximize expected social welfare subject to capacity constraint

$$\max_{t(g, w; D_N)} \mathbb{E}_\pi \left[ \sum_{(g, w)} \psi_g \theta^*(g, w) t(g, w; D_N) P(g, w) \right]$$

s.t. $\sum_{(g, w)} t(g, w; D_N) P(g, w) \leq C$ w/ probability one.

Solution is **first-best admissions rule**.
**Proposition**: First-best admissions rule is *threshold rule* w/ group-specific admissions thresholds

\[ 1 \left\{ \mathbb{E}_{\pi|D_N} [Y^* | G = g, W = w] > \tau^*(g; C) \right\}, \]

w/ ties handled s.t. capacity constraint binds.
First-best admissions rule

- **Proposition**: First-best admissions rule is *threshold rule* w/ group-specific admissions thresholds

\[
1 \left\{ \mathbb{E}_{\pi|D_N} [Y^* | G = g, W = w] > \tau^*(g; C) \right\},
\]

w/ ties handled s.t. capacity constraint binds.

- Social planner uses all available info in training data + prior to construct rank-ordering of population.
  - **Intuition**: Provided data informative about outcome, social planner “uses” data to update prior beliefs.  
  - **Intuition**: Construct optimal prediction of measured outcome \( \tilde{Y} \). Use prior beliefs \( \pi \) to map into predictions of outcome of interest \( Y^* \).
First-best admissions rule

**Proposition:** First-best admissions rule is *threshold rule* w/ group-specific admissions thresholds

\[ 1 \{ \mathbb{E}_{\pi|D_N} [Y^* | G = g, W = w] > \tau^*(g; C) \}, \]

w/ ties handled s.t. capacity constraint binds.

**Takeaways:** Deliver best estimate of \( \mathbb{E}[\tilde{Y} | W, G] \).

1. Do not blind algorithm to group membership \( G \).
2. Do not remove any characteristics \( W \).
3. Do not place additional constraints in training procedure.

Equity preferences modify admissions rule, not prediction function.
First-best admissions rule

- **Proposition**: First-best admissions rule is *threshold rule* w/ group-specific admissions thresholds

\[
1 \left\{ \mathbb{E}_{\pi|D_N} [Y^* | G = g, W = w] > \tau^*(g; C) \right\},
\]

w/ ties handled s.t. capacity constraint binds.

- **Empirical Question**: Is it really reasonable to assume social planner able to specify full prior over conditional joint distribution \((Y^*, \tilde{Y}) | W, X)\? 

- **Modelling Question**: Model assumes social planner is fully Bayesian; what if instead she is ambiguity-averse (i.e., \(\Gamma\)-minimax)?
  - Are there decision-theoretic justifications for fairness criteria?
How would social planner regulate decision-makers?

- **In many applications**: third-party firms control construction of prediction function and admissions rule
  - Examples: resume screening, credit approvals.
  Some firms may wish to discriminate against protected group.

- This is a **regulation problem**.
  - Social planner interacts w/ third-party DM. Takes their screening decisions as given.
  - Limited policy instruments to influence DM’s choices.

- **Next**: extend model to analyze second-best problem.
How would social planner regulate decision-makers?

- **In many applications**: third-party firms control construction of prediction function and admissions rule
  - **Examples**: resume screening, credit approvals.
  Some firms may wish to discriminate against protected group.

- **This is a regulation problem**.
  - Social planner interacts w/ third-party DM. Takes their screening decisions as given.
  - Limited policy instruments to influence DM’s choices.

- **Bigger Picture**: Exciting area w/ room for fruitful collaborations b/w economics and computer science.
  - How to tractably incorporate *both* richer statistics *and* richer economics into the model?
How would social planner regulate decision-makers?

- Social planner oversees market of human DMs and each faces own screening decision.

- Two constraints:
  1. **Policy constraint**: May only enforce *model regulations* – restrict which $W$ can be used in admissions rules.
  2. **Information constraint**: Does not know which human DMs are discriminatory and knows less about which $W$ have signal for predicting $Y^*$. 

- Two sets of results:
  1. Model captures existing intuitions about regulating discrimination – e.g., disparate treatment and disparate impact.
  2. Analyze how algorithmic decision-making changes this regulation problem.
Market of human decision-makers

- Human DM’s preferences $\lambda = (\lambda_0, \lambda_1)$ governs payoffs

$$U(t; \lambda) = \sum_{(g, w)} \lambda_g \theta^*(g, w) t(g, w) P(g, w),$$

Specifies relative weights placed on outcomes associated with each group.

- Only two types of preferences in the market:
  1. **Non-discriminatory** preferences w/ $\lambda_0 = \lambda_1 = 1$,
  2. **Discriminatory** preferences w/ $\bar{\lambda}_0 = 1 > \bar{\lambda}_1$

- Discriminatory firms are **taste-based discriminators** in the spirit of Becker (1957).
  - How to incorporate richer models of discriminatory behavior? E.g., stereotypes as in Bordalo et al. (2016)
Market of human decision-makers

- Human DMs prior beliefs $\pi_m$ describe beliefs about which $W$ are relevant in predicting $Y^*$.

- Each prior $\pi_m$ assoc. w/ model $m \subseteq \{1, \ldots, J\}$.
  - Only characteristics in model $m$ contain signal for predicting outcome of interest $Y^*$.

- Assume each prior $\pi_m$ additionally satisfies:
  1. **Sufficiency**: No average group differences in $Y^*$ conditional on characteristics in model $m$.
  2. **Relevance**: All characteristics in model $m$ contain signal on average for $Y^*$.
Market of human decision-makers

- Each human DM faces **capacity constraint** $C \in [0, 1]$. May not admit more than fraction $C$ of population into program.

- Market of human DMs characterized by joint dist’n $\eta(\lambda, \pi_m, C)$.
  - Full support and assume capacity constraint is independent of preferences and beliefs under $\eta$.

- Selects admissions rule to maximize expected payoffs subject to capacity constraint. Optimal admissions rule is **threshold rule**.
  - Thresholds depend on preferences $\lambda$. 
The social welfare function

- For single screening problem, SWF defined as before.

- Assume preferences are **aligned** w/ non-discriminatory human DMs.
  - Social planner’s preferred rank-ordering is same non-discriminator’s preferred rank-ordering w/ \((\psi_0, \psi_1) = (1, 1)\).

- Social planner only knows joint dist’n \(\eta\) of \((\lambda, \pi_m, C)\). Payoffs across market summarized by **aggregate social welfare function**

\[
\int_C \left( \sum_{(g, w)} \mathbb{E}_{\eta} \left[ \theta_{\pi_m}^*(w) t(g, w) \right] P(g, w) \right) h(C) dC.
\]
Model regulations

- Only policy instrument available is **model regulations**.
  - Social planner may regulate what characteristics can be used in admissions rules.

- Banning characteristics $\implies$ select rank-ordering more closely matches social planner’s preferred rank-ordering.
  - At model controls $m$, must pool across all characteristics outside of model $m$.
  - If group membership banned, further pool across groups.

Are model regulations actually enforceable?

- **Interpretation**: admissions rules are directly observable.
- **Counterpoint**: For human decision-makers, admissions rules are mechanically unobservable!
Model regulations

- Only policy instrument available is **model regulations**.
  - Social planner may regulate what characteristics can be used in admissions rules.

- Banning characteristics $\implies$ select rank-ordering more closely matches social planner’s preferred rank-ordering.
  - At model controls $m$, must pool across all characteristics outside of model $m$.
  - If group membership banned, further pool across groups.

- Are model regulations actually enforceable?
  - **Interpretation**: admissions rules are directly observable.
  - **Counterpoint**: For human decision-makers, admissions rules are mechanically unobservable!
Model regulation in practice

- Social planner can enforce model regulations $\implies$ social planner directly observes admissions rules $t(g, w)$

- **In practice**: social planner faces additional inference problem in regulating discrimination.
  - Must infer admissions rules from finitely many admissions decisions.
Model regulation in practice

- Social planner can enforce model regulations $\implies$ social planner directly observes admissions rules $t(g, w)$

- Inferring whether admissions rule uses group membership $\iff$ testing whether admissions decisions are conditionally independent of group membership.

- Conditional independence testing is a **hard** statistical problem.
  - For discrete distributions, hardness grows with dimensionality.
    Canonne et al. (2018): Tests with good size/power properties require number of samples to grow exponentially in dimensionality of observable characteristics.

- How to incorporate statistical problems into framework of optimal model regulation?
The social planner’s second-best problem

- **Second-best problem**: Select model regulations to max aggregate social welfare, taking admissions rules chosen by human DMs as given

\[
\arg \max_{m \subseteq \{1, \ldots, J\}} \int_C \left( \sum_{(g, w)} \psi_g \mathbb{E}_{\eta} \left[ \theta^*_{\pi \tilde{m}}(w) t^{\tilde{m}}_{\lambda, C}(g, w; m) P(g, w) \right] \right) h(C) dC.
\]

- Social planner searches over model controls to find regulations that induce rank-ordering most closely aligned w/ social planner’s first-best rank-ordering.

- Optimal model regulations may be quite complex.
  - Depends on fractions of discriminators vs. non-discriminators as well as dist’n of beliefs $\pi_m$
Flexibility tradeoff in model regulations

- Non-discriminators only use characteristics in admissions rule if they are predictive of the outcome $Y^*$

- Discriminators additionally may use characteristics to screen out disadvantaged group.

**Flexibility tradeoff**: Letting human DMs use additional characteristic leads to two effects

  1. Improves rank-orderings of population
  2. Used by some discriminators to screen out disadvantaged group

- Optimal model regulations involve **disparate impact tests**.
  - Does this variable provide sufficient predictive power for $Y^*$ across market relative to its predictive power of group membership $G$?
Algorithmic Decision-Making and Second-Best Model Regulations

- Considered social planner’s second-best problem when faced w/ market of human DMs.

- **Now**: introduce **algorithmic decision-making**. How does it change second-best model regulations?
  - Model algorithms as revealing *ground truth* $\theta^*(g, w)$ in each screening problem.
  - **Ground-truth model $m$**
  - **Assumption**: Firms cannot manipulate ground-truth model.

- Depends on **disclosure regime**: what must human DMs disclose about their algorithms?
  1. Only admissions rules $t(g, w)$.
  2. Both admissions rule and all model inputs (training data, training procedures, etc.).
Second-best model regulations with known admissions rules

- **First disclosure regime**: only disclose admissions rule to social planner.

- **Result**: Nothing fundamentally changes! Social planner still faces flexibility tradeoff.
  - **Intuition**: social planner still faces asymmetric information over both ground-truth model and preferences.

- **In practice**: Algorithmic decision-making forces human DMs to specify an admissions rule.
  - No longer need to infer admissions rule from finitely many admissions decisions (Kleinberg et al., 2018b).
  - Optimal regulation of admissions rules is now *feasible*. 
Algorithmic decision-making introduces new policy tool – **algorithmic audits**

- Refers to access of underlying training data and training procedures used to construct algorithm (Kleinberg et al., 2018b).

Algorithmic audits reveal ground-truth model $\theta^*_m$ of each human decision-maker.

- Eliminate one dimension of private information.

Social planner may condition model regulations on ground-truth model.

$$
\arg\max_{\tilde{m} \subseteq \{1, \ldots, J\}} \int_{C} \left( \sum_{(g, w)} \psi_g \mathbb{E}_{\lambda|m} \left[ \theta^*_m(w) t^m_{\lambda, C}(g, w; \tilde{m}) P(g, w) \right] \right) h(C) dC.
$$

**Algorithmic second-best problem.**
Second-best model regulations with algorithmic audits

- **Proposition**: Social planner allows human DMs to use any characteristics predictive of $Y^*$ at revealed ground-truth model.

- Social planner knows ground-truth model $\implies$ understand *why* a characteristic is included in admissions rule.
  - If characteristic used in admissions rule but not predictive of $Y^*$, then must be to screen out disadvantaged group!

- Requires presence of **algorithmic audits**.
Conclusion

- **This paper**: Developed economic model of screening decisions and embedding concerns about algorithmic bias within a social welfare function.

- **First-best problem**: social planner constructs prediction function and selects admissions rule.
  - *Equity irrelevance result*: equity preferences alter admissions rule, not prediction function.

- **Second-best/regulation problem**: possibly discriminatory human DMs construct prediction function and admissions rule.
  - *Algorithmic audits*: optimal to let human DMs use any characteristics that are predictive of the outcome of interest.
Conclusion

- This paper: Developed economic model of screening decisions and embedding concerns about algorithmic bias within a social welfare function.

- Optimal regulation of algorithmic decision rules is a ripe area for research and policy.
  - Analysis of the algorithmic regulation problem is area for fruitful collaboration b/w economics and computer science.
  - Several dimensions along which to enrich both the statistics and the economics of the model.
Interpreting the first-best admissions rule

- Training data **ignored** if
  \[ \mathbb{E}_{\pi|D_N}\theta^*(g, w) = \mathbb{E}_{\pi}\theta^*(g, w) \]
  for all \((g, w) \in \{0, 1\}^{J+1}\) and training datasets \(D_n\) that occur \(w/\) pos. probability.

- **Proposition**: social planner ignores \(D_N\) iff “\(Y^* \perp \tilde{Y} | W, G\)” under her prior beliefs \(\pi\).
  - Application of results in Poirier (1998) on Bayesian inference in partially identified models.

- If \(\tilde{Y}\) related to \(Y^*\) in **any way** under prior beliefs, then \(D_N\) is useful in screening decisions!
  - \(\tilde{Y}\) is mis-measured, negatively correlated \(w/\) \(Y^*\), positively correlated \(w/\) \(Y^*\) or biased against protected group!
Interpreting the first-best admissions rule

- **Another interpretation**: Construct optimal prediction of measured outcome $\tilde{Y}$ and then use prior beliefs $\pi$ to map into predictions of outcome of interest $Y^*$.
  - Extend results in Moon and Schorfheide (2012) to formalize statement.

- **Notation**: Let $\pi(\theta^*|\tilde{\theta})$ denote social planner’s conditional beliefs about $\theta^*$ given $\tilde{\theta}$.

- **Proposition**: Suppose $\hat{f}_N$ is a consistent prediction function for $\mathbb{E}[\tilde{Y} | G, W]$.
  
  Social planner’s plug-in posterior beliefs $\pi(\theta^*|\hat{f}_N)$ asymp. approx. social planner’s true posterior beliefs $\pi(\theta^*|D_N)$

  $$d_{TV}\left(\pi(\theta^*|D_n), \pi(\theta^*|\hat{f}_N)\right) \xrightarrow{p} 0 \text{ as } N \to \infty.$$
Human decision-makers: prior beliefs

- **Formal**: Joint dist’n over parameters \( \{\theta^*(g, w) : (g, w) \in \{0, 1\}^{J+1}\} \) satisfying

\[
\mathbb{E}_{\pi_m}[\theta^*(g, w_m, w_{-m})] = \mathbb{E}_{\pi_m}[\theta^*(g, w_m, w'_{-m})]
\]

for all \( g \in \{0, 1\}, w_m \in \{0, 1\}^{|m|}, w_{-m}, w'_{-m} \in \{0, 1\}^{J-|m|}. \)

- **Sufficiency**: At \( m \subseteq \{1, \ldots, J\} \) and associated beliefs \( \pi_m \),

\[
\theta^*_{\pi_m}(0, w_m) = \theta^*_{\pi_m}(1, w_m) \quad \forall \ w_m \in \{0, 1\}^{|m|}.
\]

- **Relevance**: At \( m \subseteq \{1, \ldots, J\} \) and associated beliefs \( \pi_m \),

\[
\theta^*_{\pi_m}(g, w_m) \neq \theta^*_{\pi_m}(g, w'_m) \quad w_m, w'_m \in \{0, 1\}^{|m|} \ w_m \neq w'_m.
\]
Model regulations

- **Model regulations**: Social planner may regulate what characteristics can be used in the human DMs’ admissions rules.

- If social planner implements **model regulations** $m$, then all admissions rules must satisfy

  \[ t(g, w_m, w_{-m}) = t(g, w_m, w'_{-m}) \]

  for all $g \in \{0, 1\}$, $w_m \in \{0, 1\}^{\lvert m \rvert}$ and $w_{-m}, w'_{-m} \in \{0, 1\}^{J-\lvert m \rvert}$.

- If group membership banned, then admissions rules must satisfy

  \[ t(g, w_m, w_{-m}) = t(g', w_m, w'_{-m}) \], for all $g, g'$.
Disadvantage condition

- **Disadvantage condition**: characteristics associated w/ lower avg. values of $Y^*$ more likely to occur among protected group.

- At each beliefs $\pi_m$, if $w, w'$ are s.t. $\theta^*_{\pi_m}(w) \geq \theta^*_{\pi_m}(w')$, then

  $$\frac{P(0, w)}{P(1, w)} \geq \frac{P(0, w')}{P(1, w')}.$$ 

  Holds w/ strict inequality if $\theta^*_{\pi_m}(w) > \theta^*_{\pi_m}(w')$.

- Equivalent to dist’n for protected group is *likelihood ratio dominated* by dist’n of rest of population.


