Algorithmic Classification: Theory, Data, and Welfare

Maggie Penn & John Patty
Emory University

Algorithms are used to guide high-stakes decisions about people

- Patients to treat
- Applicants approved for a loan
- Defendants that are granted bail
- Students admitted to a college
- Tax filers that are audited
- Communities police are deployed to

While algorithms may be opaque, people understand they're being classified, and may change their behavior in costly ways to obtain a good classification outcome

- Prospect of audit makes tax filers less likely to cheat
- Prospect of standardized test makes student more likely to study
- Prospect of good credit score drives responsible financial choices

These behavioral incentives may differ by group

- If I understand that it's very unlikely that a woman will be hired for a job even if qualified for it, I (as a woman) will have less of an incentive to exert costly effort to obtain qualification

This project

- An algorithm with a **general objective function** is designed to **classify a group of people**
  - Objectives can be with respect to both the **behavior** people engage in and **how they are classified**
  - *Accuracy maximization, compliance maximization, revenue maximization, hiring qualified workers, etc.*
- People want to obtain a **good classification** outcome, and can engage in a **behavior** ("compliance") to obtain a better outcome
- The algorithm is designed to **maximize its objective, knowing that people will respond to it** (a Stackelberg game)

A few takeaways on algorithms, keeping the EITM paradigm in mind

- Most work on algorithmic fairness focuses on the **statistical accuracy** of classifiers
  - Without a **theory of individual incentives and behavior**, these statistical fairness measures can be grossly misleading
  - While considered a gold standard in classification, we show that **accuracy maximization can drive inequality** across groups
- The link between between **algorithmic objectives** & **welfare** is not direct, though algorithms are often described in normative terms
  - Increasing algorithm's "taste for punishment" (making it more predatory) can be **Pareto improving**
- **EITM takeaway**: *Using data to make normative judgments about human outcomes requires an explicit theory of what people want!*

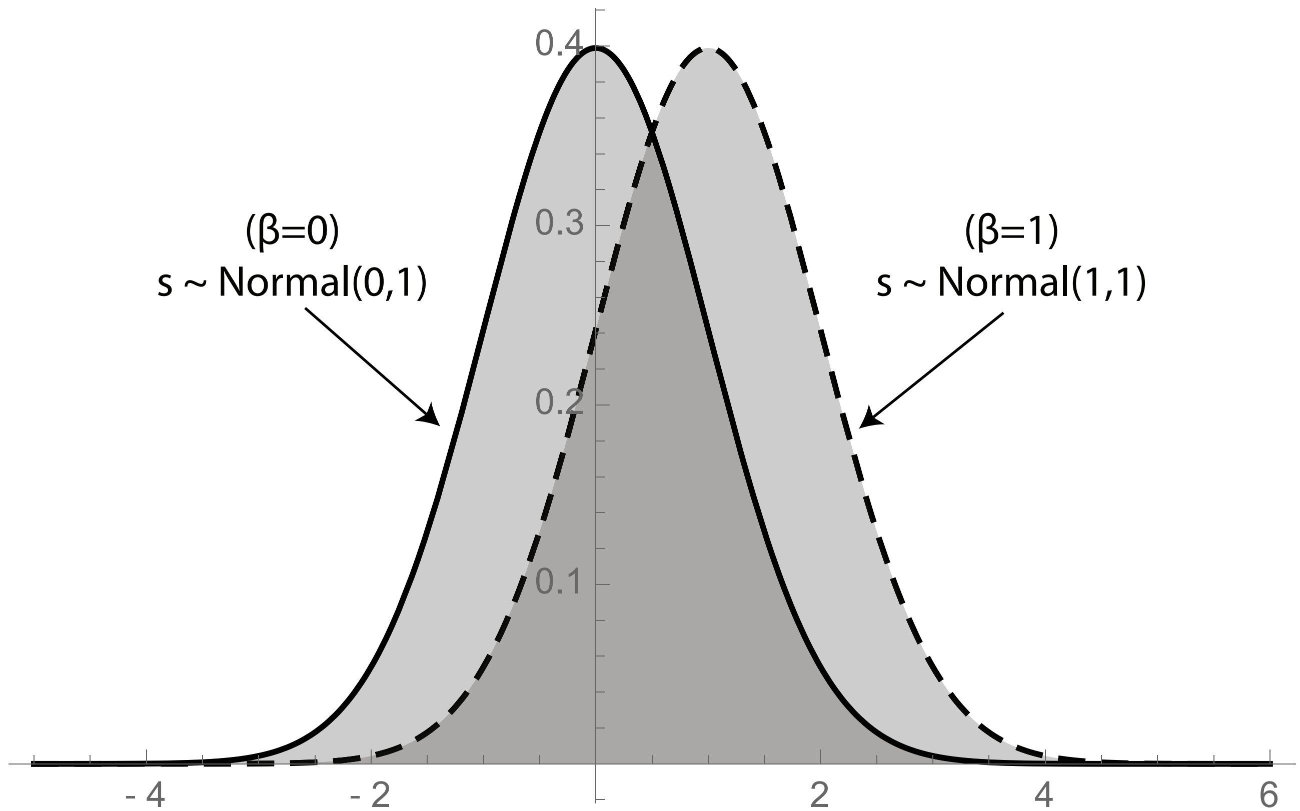
The Model (Individuals)

- A unit mass of **individuals** N, with $i \in N$
- Each person chooses a costly **behavior** $\beta_i \in \{0, 1\}$ ("compliance")
  - $\beta_i \in B$ represents an activity that each person will be **classified**, and **potentially rewarded**, on the basis of
- Person i pays **private cost** γ_i to choose $\beta_i = 1$
  - Costs γ distributed with CDF F
  - **Example**: β_i captures "lawfulness" and γ_i is i's "cost to being lawful"

The Algorithm

- Algorithm observes **signal** s_i about behavior β_i drawn from behavior-dependent PDF g_β (g_1 and g_0 satisfy the MLRP)
  - The higher the signal, the more likely it was that the person complied
- Signal could be a unidimensional **test result**
- Or we can think of each person's set of covariates $x_i \in X$ as associated with a **likelihood ratio** that *is* the signal

$$s_i = \frac{P(x_i|\beta_i = 1)}{P(x_i|\beta_i = 0)}$$



Example of signal distributions conditional on behavior β

Classification

- After observing signal s the algorithm makes a **binary classification decision** for each person, $d_i \in \{0, 1\}$

- The algorithm's **strategy** δ maps each signal s_i into a probability of reward:

$$\delta(s) = \Pr[d_i = 1 | s_i]$$

- If $d_i = 1$ then i gets a **reward**, if $d_i = 0$ then i pays a **penalty**

Individual Payoffs

- Each person receives the following payoff:

$$u(\beta_i, d_i | \gamma_i) = \underbrace{r \cdot d_i}_{\text{bonus if classified 1}} - \underbrace{\gamma_i \cdot \beta_i}_{\text{cost if compliant}}$$

- $r > 0$ is an exogenous parameter capturing the **"stakes" to classification**

$$r = (\text{reward if classified } d = 1) - (\text{penalty if classified } d = 0)$$

$\Rightarrow$ People benefit from receiving a positive classification

Individual Behavior

- The individual's **incentives** $\Delta(\delta)$ capture the net benefit to any person of choosing $\beta_i = 1$ over $\beta_i = 0$

$$\Delta(\delta) = r \int_{s \in \mathbf{R}} (g_1(s) - g_0(s)) \cdot \delta(s) ds$$

- **A person chooses** $\beta_i = 1$ **if**:

$$\underbrace{\gamma_i \leq \Delta(\delta)}_{\text{cost to compliance low enough}}$$

- Algorithm is "**behaviorally null**" if $\Delta(\delta) = 0$
  - If stakes to classification $r = 0$
  - If $\delta(s) = c$ for all s (algorithm classifies everyone the same way)

The Algorithm's Objectives

	Decision	
Behavior	$d_i = 1$	$d_i = 0$
$\beta_i = 1$	A_1 (True Positive)	A_0 (False Negative)
$\beta_i = 0$	B_0 (False Positive)	B_1 (True Negative)

Algorithm receives "payoff" $A_1, A_0, B_1, B_0 \in \mathbf{R}$ for % of people that fall into each cell

Algorithm optimally designed to **generate behavior** (β) and **bin signals of behavior** (d) into most beneficial cells of matrix

Two Examples of Algorithm Objectives

Accuracy

	$d_i = 1$	$d_i = 0$
$\beta_i = 1$	$A_1 = 1$	$A_0 = 0$
$\beta_i = 0$	$B_0 = 0$	$B_1 = 1$

Compliance

	$d_i = 1$	$d_i = 0$
$\beta_i = 1$	$A_1 = 1$	$A_0 = 1$
$\beta_i = 0$	$B_0 = 0$	$B_1 = 0$

Timing of Decisions

1. People **privately observe their costs** to compliance, γ_i
2. An algorithm $\delta(s)$ is publicly **chosen / committed to**
   - Algorithm knows cost distribution F, signal distributions g_β
3. People make their **compliance decisions** $\beta_i \in \{0, 1\}$
4. **Signals are generated** and **classified** according to algorithm $\delta(s)$
5. **Payoffs are distributed** to people and the algorithm

Optimal classifiers have a simple characterization

- The "best" algorithm sets a cutpoint $\tau^* \in \mathbf{R} \cup \pm\infty$ and utilizes either a **threshold** or **negative threshold** rule

Threshold rule

$$\overline{\tau}^*(s_i) = \begin{cases} 1 & \text{if } s_i \geq \overline{\tau}^* \\ 0 & \text{otherwise.} \end{cases}$$

Negative threshold rule

$$\underline{\tau}^*(s_i) = \begin{cases} 0 & \text{if } s_i \geq \underline{\tau}^* \\ 1 & \text{otherwise.} \end{cases}$$

How can negative threshold rules be optimal?

- These rules punish people with signals *above* some threshold, so those more likely to have complied are punished

- **This disincentivizes compliance**

  - Designer might have a direct taste for non-compliant behavior

  - *Or* inducing non-compliance might make other goals (e.g. accuracy!) easier to achieve

- A negative threshold is "cheapest" way to induce non-compliance

  - Provides **greatest behavioral incentive** to not comply (MLRP)

  - **Fewest misclassifications** in the tail of the signal distribution

Example: Accuracy maximization drives inequality

- Consider two groups that differ in their members' average costs to compliance
- **Low cost group** has costs distributed $N[\frac{1}{2}, 1]$
  - $\mathbf{31\%}$ **of the population complies** without any extrinsic incentives
- **High cost group** has costs distributed $N[\frac{3}{4}, 1]$
  - $\mathbf{23\%}$ **of the population complies** without any extrinsic incentives

Most Accurate Classifiers for the Two Groups

- For the **low-cost** group, the most accurate classifier is a positive threshold rule with $\overline{\tau}^* \approx -0.2$
  - Equilibrium compliance is 85% (**increased** from 31%)
  - **This classifier is 81% accurate**
- For the **high-cost** group, the most accurate classifier is a negative threshold rule with $\underline{\tau}^* \approx -1.4$
  - Equilibrium compliance is 13% (**decreased** from 23%)
  - **This classifier is 80% accurate**

Stakes $r = 5$ and signal distributions are $g_0 = N[0, 1]$ and $g_1 = N[1, 1]$

Takeaways about “most accurate” algorithms

- Accuracy motivations often thought of as **fair** or **neutral**
  - The algorithmic fairness literature focuses largely on statistical error rates across groups in classification outcomes
  - Here, both groups are correctly classified $\approx 80\%$ of the time
  - But the algorithm incentivizes **opposite** behavior for the groups, exacerbating a kind of societal/behavioral inequality across the groups
- Because data and behavior are **performative** (respond to the algorithm), accuracy-maximization entails manipulating behavior to overcome noisy data

How robust is this example?

Proposition

For *any* reward $r > 0$ and *any* signal accuracy we can find two cost distributions F_X and F_Y for which the accuracy maximizing designer

- **strictly incentivizes compliance** for Group X and
- **strictly incentivizes non-compliance** for Group Y

Example: Algorithm objectives and social welfare

Accuracy

	$d_i = 1$	$d_i = 0$
$\beta_i = 1$	$A_1 = 1$	$A_0 = 0$
$\beta_i = 0$	$B_0 = 0$	$B_1 = 1$

Accurate + predatory

	$d_i = 1$	$d_i = 0$
$\beta_i = 1$	$A_1 = 1$	$A_0 = 0.5$
$\beta_i = 0$	$B_0 = 0$	$B_1 = 1.5$

- Most accurate algorithm sets $\overline{\tau}^* = \mathbf{0.462}$, yields 98% compliance
- "Predatory" algorithm sets $\overline{\tau}^* = \mathbf{0.125}$, yields 96% compliance
- The predatory algorithm is more lenient, and is ex ante preferred to the most accurate algorithm **by every person being classified**

Costs $\gamma \sim N[0, 1]$, stakes $r = 2$, signals $g_0(s) = N[0, 0.1]$, $g_1(s) = N[1, 0.1]$.

Takeaways: Algorithm objectives and social welfare

- Without a theory of individual preferences, there's **no reason to think accurate classification is desirable** from standpoint of those classified

- (More provocatively?) we should be careful using an **algorithm's objectives to make welfare judgments** without a theory of behavior
  - In this case, directly increasing the algorithm's "taste" for punishment (giving it a payoff bump for every $d = 0$) **strictly benefits every person** in expectation

Conclusions

- The prospect of being classified affects the **life choices people make**

- When data are **performative** (respond to how the data are used, as is often true of data about *people*) we need a theory of the data generating process to make normative judgments

- Fairness with respect to how **data are used** (e.g. statistical accuracy) might be at odds with fairness in the **data generating process**

- *EITM can help us make sense of these tensions!*

Related Literature

- **Algorithmic Fairness & Welfare** (Hu & Chen, 2019; Liang & Lu, 2024)
  - Welfare effects of fair classification with fixed outcomes
- **Behavioral Effects of Classification Design**
  (Jung, *et al*, 2020; Coate & Loury, 1993)
  - Theoretically proximate; Jung considers compliance maximization; Coate models a simultaneous move game with multiple equilibria
- **Strategic Classification / Performative Prediction**
  (Hardt, *et al*, 2016; Hu, *et al* 2018; Perdomo, *et al*, 2021)
  - Classification with endogenous (observable) data
- **Outcome Performativity** (Kim & Perdomo, 2023)
  - Classification with endogenous *outcomes*
  - Focus is whether data/outcomes can be <u>learned</u>; we assume alg *knows* how data respond to it (our focus is behavior & welfare)