NBER WORKING PAPER SERIES

AN ECONOMIC APPROACH TO REGULATING ALGORITHMS

Ashesh Rambachan
Jon Kleinberg
Sendhil Mullainathan
Jens Ludwig

Working Paper 27111
http://www.nber.org/papers/w27111

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2020, Revised January 2021

We thank Alex Frankel, Ed Glaeser, Jonathan Guryan, Robert Minton, Joshua Schwartzstein, participants at the Labor/Public Breakfast at Harvard University, the Seminar in Law, Economics & Organization at Harvard Law School, AEA Session on Algorithmic Fairness and Bias and the NBER Conference on the Economics of AI for valuable feedback. We are especially grateful to Joshua Gans, Paul Milgrom and Hal Varian for their constructive comments and feedback as conference discussants. Rambachan gratefully acknowledges financial support from the NSF Graduate Research Fellowship (Grant DGE1745303). All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

An Economic Approach to Regulating Algorithms
Ashesh Rambachan, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig
NBER Working Paper No. 27111
May 2020, Revised January 2021
JEL No. C54,D6,J7,K00

ABSTRACT

There is growing concern about "algorithmic bias" - that predictive algorithms used in decision-making might bake in or exacerbate discrimination in society. We argue that such concerns are naturally addressed using the tools of welfare economics. This approach overturns prevailing wisdom about the remedies for algorithmic bias. First, when a social planner builds the algorithm herself, her equity preference has no effect on the training procedure. So long as the data, however biased, contain signal, they will be used and the learning algorithm will be the same. Equity preferences alone provide no reason to alter how information is extracted from data - only how that information enters decision-making. Second, when private (possibly discriminatory) actors are the ones building algorithms, optimal regulation involves algorithmic disclosure but otherwise no restriction on training procedures. Under such disclosure, the use of algorithms strictly reduces the extent of discrimination relative to a world in which humans make all the decisions.

Ashesh Rambachan
Department of Economics
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
asheshr@g.harvard.edu

Jon Kleinberg
Department of Computer Science
Department of Information Science
Cornell University
Ithaca, NY 14853
kleinber@cs.cornell.edu

Sendhil Mullainathan
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
Sendhil.Mullainathan@chicagobooth.edu

Jens Ludwig
Harris School of Public Policy
University of Chicago
1307 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

1 Introduction

The growing use of algorithms to inform consequential decisions such as hiring, credit approvals, pre-trial release and medical testing has been accompanied by growing concerns of "algorithmic bias." Lacking a single definition, the blanket term algorithmic bias is often used to describe fears that algorithmic decision-making may exacerbate existing discrimination and inequality.[1] These fears arise in part because the data used to train algorithms often reflect historical discrimination and inequality. For example, resume screening software is trained upon past hiring decisions, which themselves might have been discriminatory. Criminal records used in recidivism prediction may bake in differential arrest rates by police or conviction rates by judges. A theoretical literature, largely in computer science, explores how such biases arise and how their presence should change the design and use of algorithms.

Concerns about algorithmic bias, at their heart, are questions of optimal policy and existing research typically misses three ingredients that economists view as central to any policy analysis. First, there is rarely a clear specification of the social planner's preferences over outcomes (and how they are distributed across the population) from which optimal policy can be derived. Instead, in most existing work, fairness is often defined as a property of the algorithm that is imposed as an additional constraint on the training procedure, rather than being derived from the outcomes produced by the resulting decisions. Second, there is rarely a description of the policy tools available that could influence outcomes beyond constraining the algorithm itself. Finally, incentive problems are often overlooked; instead it is assumed that the algorithm designer shares society's preferences.

In this paper, we incorporate the new issues raised by the use of supervised machine learning algorithms in decision-making into a canonical welfare economics framework that includes these three ingredients. We focus on situations in which an empirically-based supervised machine learning algorithm informs a *screening decision* - a situation where one or more people must be selected from a larger pool based upon a prediction of an uncertain outcome of interest for each of them. We model a supervised machine learning algorithm as consisting of two components: a "predictive algorithm," which takes in training data consisting of outcomes and observed characteristics for a set of individuals and returns a prediction function, and a "decision rule," which uses the constructed prediction function to make decisions. A policymaker therefore has two distinct tools

[1] The literature on algorithmic bias is vast. Barocas et al. (2019) is a textbook introduction to the computer science literature on this topic. Chouldechova and Roth (2020) provides a recent overview of this literature as well. Cowgill and Tucker (2019) provides a survey for economists.

available: influencing the design of the predictive algorithm and influencing how the decision rule uses the predictions. The social welfare function is defined over the resulting outcomes of the screening problem, incorporating both a concern for efficiency and an explicit preference for more equitable outcomes across groups. We take the social welfare function as a primitive and derive its implications for algorithm construction and regulation in two policy environments.

We first study the algorithm choice of a social planner who wishes to maximize social welfare and makes the screening decisions herself - the "first-best problem." We find an *equity irrelevance* result: the planner's equity preferences have no effect on how the predictive algorithm is constructed. All characteristics, including group membership, are given to the predictive algorithm, and fairness concerns do not lead the social planner to place any additional constraints on her training procedure. Since the predictive algorithm simply summarizes information in the observed training data, the social planner does not wish to destroy potentially useful information no matter her preferences. This is robust to a wide variety of common concerns surrounding sources of algorithmic bias. It holds even if the observed outcome in the training data differs from the outcome of interest in some way that is systematically related to group membership, or if there are differences in the conditional base rates of the outcome of interest across groups, or if there are differences in the distribution of characteristics across groups.

As an illustrative example, consider an employer with a fixed number of job openings available and a large pool of applicants that apply to the job openings. Each applicant is described by observed characteristics that are gleaned from their resume. Applicants fall into two groups: an advantaged group (e.g., whites, males) and a disadvantaged group (e.g., Blacks, females). The employer must make predictions about each applicant's on-the-job productivity based on their observed characteristics. In doing so, the employer has access to a historical dataset that consists of the resumes of its employees along with some measure of their on-the-job performance, such as performance reviews. The employer may attempt to learn the relationship between the information available in resumes and whatever available measure of productivity using a supervised machine learning algorithm. In this example, our first best analysis focuses on the case in which the employer shares society's goals and wishes to hire more members from the disadvantaged group. Our equity irrelevance result establishes that possible biases in the historical data, for example biased performance reviews, do not change the benevolent employer's desire to use the algorithm. Indeed so long as the data, as biased as they are, include *any* signal that is relevant for underlying worker productivity, the employer will wish to use these data.

Next, we consider the "second best" *regulation problem*, where the social planner neither makes the screening decisions nor the algorithm design choices; instead private actors, some with discriminatory preferences, do so. For example, in hiring, individual firms decide who to hire and how to do that hiring. Therefore producing equitable employment outcomes involves implementing regulations around the hiring process. Our model of this regulation problem analyzes an environment in which some private actors are taste-based discriminators in the spirit of Becker (1957), there are no average group differences in the outcome of interest conditional on observable characteristics and that the social planner has no further equity preferences beyond limiting discrimination (i.e., no explicit affirmative action motive).

We begin by analyzing optimal regulation absent algorithms in which the regulator must decide what kinds of characteristics a human decision-maker can use in their decision (e.g. hiring) rules. The model captures many of our existing intuitions about the difficulty of regulating discrimination and we show that optimal regulation resembles existing anti-discrimination policy: prohibitions against disparate treatment (direct use of protected characteristics to treat individuals differently) as well as tests for disparate impact (use of ostensibly neutral factors that wind up disproportionately harming one group). Intuitively, the social planner faces a "flexibility tradeoff" in regulation – allowing more characteristics to be used leads to more accurate predictions, but the flexibility to use extra characteristics may also be used to screen out members of the disadvantaged group. The equilibrium level of discrimination is strictly positive in this market, meaning that the flexibility tradeoff poses a fundamental challenge to regulating discrimination.

We then consider the case in which private actors use an algorithm in their screening decisions. Optimal regulation changes substantially as long as there is full disclosure of all parts of the algorithm: the data, training procedure and decision rule. We refer to such disclosures as an *algorithmic audit* (Kleinberg et al., 2018, 2020). With such algorithmic audits in place, it is now optimal to allow the predictive algorithm to access to any characteristic that is predictive of the outcome of interest. The ability to carry out algorithmic audits enables the social planner to enforce that all human decision-makers (discriminatory or non-discriminatory) with the same prediction function select the same admissions rule. This mechanism strictly reduces the equilibrium level of discrimination relative to a world in which algorithms are not used. With the correct regulatory system in place – specifically, one that allows algorithmic audits to be conducted – the introduction of predictive algorithms into screening decisions leads to not only improved prediction, but simultaneously makes it easier to detect discrimination in the market.

Returning to the hiring example, now consider the case in which firms make their

own hiring decisions and some wish to discriminate against the disadvantaged group. The social planner may influence firms by placing restrictions on the characteristics used in hiring decisions.[2] Absent algorithms, the social planner's flexibility tradeoff summarizes the well-known difficulties of detecting discrimination in hiring decisions. The more applicant characteristics firms are able to use, the better able they are to accurately predict productivity. At the same time, discriminating firms are also more able to find characteristics that help predict group membership as well, thereby enabling these firms to screen out members of the disadvantaged group. The flexibility tradeoff arises because the social planner faces two sources of asymmetric information relative to the firms – she neither knows which firms are discriminatory nor which characteristics are predictive of productivity. In sharp contrast, if firms adopt predictive algorithms in their hiring decisions *and* there is full algorithmic disclosure, the social planner's regulation problem fundamentally changes. Under full algorithmic disclosure, the social planner no longer faces asymmetric information over which characteristics are predictive of productivity. For a set of firms that face the same relationship between applicant characteristics and productivity (i.e., firms for whom the underlying prediction function is the same), the social planner can now force discriminatory and non-discriminatory firms to make the same hiring decisions. Under such a disclosure regime, our results establish that it is optimal to let any characteristic that is predictive of the outcome of interest be used in hiring decisions and the equilibrium level of discrimination is zero.

Implicit throughout our analysis is the important assumption that there is a *consensus* on the choice of outcome. By explicitly writing a social welfare function, we are assuming that there is social consensus on the outcome of interest, yet in some settings there may be substantive disagreement. For example, in pre-trial release decisions, does social welfare only depend on pre-trial misconduct rates among the released? In criminal sentencing, does social welfare only depend on future recidivism (Doleac and Stevenson, 2019)? Similarly, in the regulation problem, we assume that the firms' payoffs depend on the same outcome as the social welfare function. Yet, is there a common definition of a "good employee" in our running example on hiring decisions? Recent empirical work has documented that documented algorithmic bias in some settings is driven precisely by disagreement over the outcome (Passi and Barocas, 2019). For example, Obermeyer et al. (2019) analyzed an algorithmic decision tool that generated large racial disparities across patients. These disparities arose because the underlying prediction function was trained

[2]This is consistent with the observation that regulators rarely dictate to firms exactly how many people to hire in practice, but regulators do tell firms how they are allowed to select those people they do hire - for example, prohibiting firms from explicitly using group membership itself as a criteria in hiring decisions.

to predict the cost of caring for the patient, which the health care provider cares about, as opposed to a measure of patient health. Understanding how disagreement over the outcome shapes the design and regulation of algorithmic decision-making is an important avenue for future research that lies beyond the scope of this paper.

Related literature: Our approach is different from that taken by a large community of researchers in computer science. Existing research typically begins by noting that a supervised machine learning algorithm generates a mapping from observed data into predictions or decisions and then formally defines what it means for these mappings to be "fair." Given a particular definition, researchers then ask how to construct such fair mappings from data and whether a given algorithm satisfies this property. This framework is enormously influential, generating numerous important insights in the study of algorithmic decision-making. Canonical papers in computer science include Dwork et al. (2012), Zemel et al. (2013), Feldman et al. (2015), Hardt et al. (2016), Corbett-Davies et al. (2017), Raghavan et al. (2017) and Chouldechova (2017).

In contrast, we define fairness in terms of preferences over the resulting outcomes of the screening decision using a social welfare function. We take the preferences summarized by the social welfare function as our primitive and derive its implications for algorithm construction. Several papers in computer science also examine the connections between definitions of predictive fairness in computer science and social welfare. See, for example, Hu and Chen (2018), Heidari et al. (2018), Balashankar et al. (2019) and Hu and Chen (2020). However, this research tends to focus on the first-best problem in which a benevolent planner controls the design and implementation of the algorithm, overlooking agency problems that arise when algorithms are designed and implemented by third-party decision-makers. Our analysis of the regulation problem is a new contribution, highlighting the value of an economic perspective. We discuss the connections to the literature in computer science in more detail in Section 3 after we establish our framework. Appendix A discusses how our framework accommodates and relates to several common worries surrounding algorithmic bias.

Finally, we highlight several recent papers in economics that also incorporate insights from economics into the study of algorithmic decision-making. Athey et al. (2020) studies the optimal delegation rule of a principal that may either delegate decision-making to an algorithmic decision-rule or a human decision-maker. Cowgill and Stevenson (2020), applying classic strategic communication models, highlights that if predictions are manipulated by a planner, then human decision-makers may optimally ignore these predictions in their decisions. Finally, our formal analysis of the regulation problem formally builds

upon ideas first discussed in Kleinberg et al. (2018, 2020).

2 The Screening Decision and the Social Welfare Function

We introduce the key building blocks of our model by defining the screening decision and the social welfare function. There is a population of individuals that are to be screened into a program based on predictions of an unknown outcome of interest. Each individual is described by a vector of observable characteristics, and these characteristics may be used to predict the outcome of interest. The social welfare function is defined in terms of the resulting outcomes of the screening decisions.

2.1 The population of individuals

There is a unit mass of individuals divided into two groups, denoted $G \in \{0,1\}$. We refer to $G = 1$ as the "disadvantaged group." Each individual in the population is described by a vector of characteristics $W := (W_1, \ldots, W_J) \in \{0,1\}^J$. Each individual is also associated with two real-valued, discrete labels $Y^* \in \{y_1^*, \ldots, y_L^*\}, \tilde{Y} \in \{\tilde{y}_1, \ldots, \tilde{y}_K\}$, where the label Y^* is the "outcome of interest" and the label $\tilde{Y}$ is the "measured outcome." Assume, without loss of generality, that the labels are ordered, meaning $y_1^* \leq \ldots \leq y_L^*$ and $\tilde{y}_1 \leq \ldots \leq \tilde{y}_K$. The population of individuals is summarized by a joint distribution $\mathbb{P}$ over the random vector $(Y^*, \tilde{Y}, G, W)$.

Let $P(g,w) := \mathbb{P}\{G = g, W = w\}, P(w) := \mathbb{P}\{W = w\}$ be the fraction of individuals that belong to group g with characteristics $w \in \{0,1\}^J$ and the fraction of individuals with characteristics w respectively. Assume that $P(g,w) > 0$ for all $(g,w) \in \{0,1\}^{J+1}$. Finally, let $\theta^*(g,w) := \mathbb{E}[Y^* \mid G = g, W = w]$, $\tilde{\theta}(g,w) := \mathbb{E}\left[\tilde{Y} \mid G = g, W = w\right]$ denote the average outcome of interest Y^* and the average measured outcome $\tilde{Y}$ among individuals that belong to group g with characteristics w. In Appendix A, we discuss in more detail how our setting nests several common sources of bias mentioned in the computer science literature on algorithmic bias.

2.2 The screening decision

Individuals in the population may be granted admission into a program. The program is capacity constrained and only a fraction $C \in [0,1]$ of the population may be granted admission. The information available when making the admissions decisions are the observed characteristics W and group membership G. A *decision rule* denoted $t(g,w) \in [0,1]$ describes the probability that an individual in group g with characteristics w is admitted

into the program. The capacity constraint implies that

$$\sum_{(g,w)\in\{0,1\}^{J+1}} t(g,w)P(g,w) \leq C. \tag{1}$$

As we will see next, the social planner would like to make the admissions decisions based on the outcome of interest Y^*. However, since Y^* is not observed for any given individual in the population at the time of the decision, the admissions decisions will instead be based upon predictions of the unknown outcome Y^*. These predictions will use the observed characteristics (G, W) and the social planner's beliefs about the joint distribution of $(Y^*, \tilde{Y}, G, W)$ in the population.

Before continuing, we introduce two simple examples to illustrate how our model maps into common screening problems of interest.

Example 1 (Hiring)**.** *Which applicants should be hired for a job? Applicants are described by a vector of characteristics (W) that may be gleaned from their submitted resumes. For example, these characteristics may include traditional information such as the applicant's education and work history. It may also include "high-dimensional" features that are parsed used natural language processing algorithms such as the frequency of certain words on the resume. Applicants have some unobserved productivity on the job (Y*) and we wish to infer their productivity from the observed resume.*

Example 2 (Loan decisions)**.** *Which individuals should be granted a loan? Individuals submit an application and other information to a financial institution for a loan. Applicants are described by a vector of characteristics (W) that are contained in the application and other financial information that is available to the financial institution. For example, this may include traditional financial information such as the applicant's reported income, outstanding debt and stated purpose of the loan. It may also include a rich set of high-dimensional, high-frequency transaction level data that the financial institution has access to if the applicant is an existing customer. Applicants have an unobserved probability of repaying the loan (Y*) and we wish to infer their probability of loan repayment from the application.*

2.3 The social welfare function

The social welfare function defines society's preferences over the outcomes produced by the screening decisions. It is a weighted average of the outcome of interest among individuals that are admitted into the program:

$$\sum_{(g,w)\in\{0,1\}^{J+1}} \psi_g \theta^*(g,w)t(g,w)P(g,w), \tag{2}$$

where $\psi_g \geq 0$ are generalized social welfare weights placed upon individuals in group g. The social welfare weights imply that the outcome of interest may be valued differently across groups. If $\psi_1 > \psi_0$, then the outcomes associated with the disadvantaged group are valued more than outcomes associated with the rest of the population, which implies that for a given average value of the outcomes among admitted people we would prefer to admit more members of the disadvantaged group, capturing a preference for "equity." Moreover, since the social welfare function is defined directly in terms of the average outcome of interest of the admitted group, it is larger whenever the admitted set has higher average values of the outcome of interest, holding fixed the fraction of the population that is admitted into the program and the composition of the admitted group. This captures a preference for more "efficient" outcomes.

In Appendix B, we provide a motivation for the social welfare function in Equation 2. We sketch a setting in which the utility of each individual depends on some measured outcome and whether they are admitted into the program. The true outcome of interest is therefore the change in the utilities of an individual from being admitted into the program at a given value of the observed outcome. The social planner's welfare weights may be higher on the disadvantaged group because, for instance, the utility of an individual from the disadvantaged group may be uniformly lower than the utility of an individual from the advantaged group. This may capture either unmodeled discrimination against the disadvantaged group or existing disparities across groups in a reduced form manner.

Example 1 (continuing from p. 8)**.** *More productive workers (higher Y^*) produce output if hired and the social welfare function depends on total output. However, the social planner wishes to protect Black workers ($G = 1$), and so places a larger weight on output produced by them in the social welfare function ($\psi_1 > \psi_0$).*

Example 2 (continuing from p. 8)**.** *Loans are given out to consumers and more credit-worthy borrowers (higher Y^*) are less likely to default on their loans. The social welfare function depends on the total loan repayment rate. In addition, the social planner wishes to ensure that minority borrowers ($G = 1$) have access to credit, and places more weight on credit access among these groups.*

2.4 The training dataset

From the social welfare function in Equation (2), it is immediate that the social planner wishes to select an admissions rule $t(g, w)$ that is based on the average outcome of interest $\theta^*(g, w)$. If $\theta^*(g, w)$ were known, the social planner would simply construct a rank-ordering of the population in terms of the welfare-weighted average outcome of interest

$\psi_g \theta^*(g, w)$, admitting individuals into the program in descending order until she reaches the capacity constraint C. However, the average outcome of interest $\theta^*(g, w)$ is unknown, and the social planner faces a non-trivial "prediction policy problem" (Kleinberg et al., 2015).

To construct estimates of $\theta^*(g, w)$, the social planner has access to a *training dataset* that consists of N randomly drawn samples from the population of individuals. For each observation in the training dataset, the characteristics (G, W) and the measured outcome $\tilde{Y}$ are recorded. Let $D_N = \{(\tilde{Y}_i, W_i, G_i)\}_{i=1}^N$ denote the observed training dataset. Even though the training dataset D_N does not contain the outcome of interest Y^*, it may still be useful in constructing predictions. For example, if the measured outcome $\tilde{Y}$ is correlated with the outcome of interest Y^*, then there may be useful information in the training dataset.

Example 1 (continuing from p. 8)**.** *We would prefer to hire workers that are productive but productivity is unobserved. Instead, among currently hired employees, we observe performance reviews $\tilde{Y}$, which is a possible proxy for productivity. A training dataset D_N may consist of the observed characteristics of past and current employees along with their performance reviews.*

Example 2 (continuing from p. 8)**.** *We would like to grant loans to applicants that will repay. Among current borrowers, we observe whether they have missed repayments or have made late payments $\tilde{Y}$, which is a possible proxy for repayment ability. A training dataset D_N may consist of past and current loans along with their repayment history.*

3 The Social Planner's First-Best Algorithm Design

In this section, we define the social planner's first-best problem and characterize its solution. The social planner is a Bayesian decision-maker, specifying her prior beliefs about the conditional joint distribution of the measured outcome $\tilde{Y}$ and the outcome of interest Y^* given the characteristics (G, W). The social planner uses the observed training dataset D_N to update these beliefs.

We then characterize the social planner's optimal algorithm that maximizes social welfare. For a fixed training dataset D_N, the social planner's optimal algorithm ranks the population using her posterior beliefs about the average outcome of interest $\theta^*(g, w)$ and then admits individuals into the program based on this ranking, applying a group-specific threshold for admission. Next, as the size of the training dataset D_N grows large, the ranking used by the social planner is equivalent to the ranking that would be produced by constructing a consistent estimate of the average measured outcome $\tilde{\theta}(g, w)$ and then applying an ex-post adjustment based on her prior beliefs about the relationship between the measured outcome and the outcome of interest. Together these results

imply a strong form of *equity irrelevance* - the social planner's equity preferences only modify the decision rule, not the predictive algorithm.

3.1 The social planner's beliefs

We assume that the social planner knows the marginal distribution of the characteristics (G, W) in the population and only faces uncertainty over the conditional joint distribution of the measured outcome $\tilde{Y}$ and the outcome of interest Y^*. The social planner is a Bayesian decision-maker with prior beliefs about how the measured outcome relates to the outcome of interest.

Formally, for outcomes $y_l^*, \tilde{y}_k$, define the parameters

$$\mathbb{P}\left\{Y^* = y_l^*, \tilde{Y} = \tilde{y}_k | G = g, W = w\right\} := \eta_{l,k}(g, w) \tag{3}$$

$$\mathbb{P}\left\{Y^* = y_l^* | G = g, W = w\right\} = \sum_{k=1}^{K} \eta_{l,k}(g, w) := \eta_l^*(g, w) \tag{4}$$

$$\mathbb{P}\left\{\tilde{Y} = \tilde{y}_k | G = g, W = w\right\} = \sum_{l=1}^{L} \eta_{l,k}(g, w) := \tilde{\eta}_k(g, w). \tag{5}$$

Let $\eta := \{\eta_{l,k}(g, w) : l \in \{1, \ldots, L\}\ k \in \{1, \ldots, K\},\ g \in \{0, 1\},\ w \in \{0, 1\}^J\}$ collect together these parameters at each possible value of the characteristics (G, W). The social planner's prior beliefs are a prior distribution $\pi(\cdot)$ over the finite dimensional parameter η.

The social planner uses the observed training data to update her prior beliefs $\pi(\cdot)$, forming a posterior distribution $\pi|D_N$. The likelihood function for the observed training data is simply

$$\mathcal{L}(D_N; \eta) := \prod_{i=1}^{N} \left(\prod_{k=1}^{K} \tilde{\eta}_k(G_i, W_i)^{1\{\tilde{Y}_i = \tilde{y}_k\}} \right) P(G_i, W_i). \tag{6}$$

Since the marginal distribution of (G, W) is known, the likelihood function only depends on the observed training dataset D_N and the parameters η but not the marginal distribution $P(g, w)$. Applying Bayes' rule, the social planner uses the observed training dataset to construct her posterior beliefs $\pi|D_N$.

3.2 Characterizing the social planner's first-best admissions rule

Given the social welfare function and her prior beliefs π, the social planner selects an admission rule to maximize expected social welfare subject to her capacity constraint $C \in [0, 1]$. This is the social planner's *first-best problem*.

Definition 1. *Given prior beliefs π, social welfare weights (ψ_0, ψ_1) and capacity constraint C, the social planner's* ***first-best problem*** *is*

$$\max_{t(g,w;D_N)} \mathbb{E}_\pi \left[\sum_{(g,w)} \psi_g \mathbb{E}_\eta \left[\theta^*(g,w) t(g,w;D_N) \right] P(g,w) \right]$$
$$s.t. \sum_{(g,w)} t(g,w;D_N) P(g,w) \leq C \quad \textit{with probability one.}$$

The solution $t^*(g,w;D_N)$ for all $(g,w) \in \{0,1\}^{J+1}$ is the social planner's ***first-best algorithm****.*

The social planner's first-best problem is to select a data-driven algorithm $t(g,w;D_N)$ to maximize expected social welfare, where the social planner uses her prior beliefs π to average over possible realizations of the training dataset and the parameter η. The capacity constraint must hold at all realizations of the training dataset that occur with positive probability.

The social planner's first-best algorithm consists of two components: a decision rule, which is a threshold rule with group-specific thresholds for admission, and a predictive algorithm that rank-orders the population based upon a prediction of the outcome of interest.

Proposition 1. *The social planner's first-best admissions rule is a threshold rule with group-specific admissions thresholds*

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

where ties with $\mathbb{E}_{\pi|D_N} \left[\theta^*(g,w) \right] = \tau^*(g;C)$ are handled such that the capacity constraint holds with equality.

The social planner uses her prior beliefs $\pi(\cdot)$ and the observed training data D_N to construct the best rank-ordering of the population in terms of the expected value of Y^* given the observed characteristics G, W. This ranking is encapsulated in her posterior beliefs $\pi|D_n$, which conditions on the entire training dataset. The social planner's optimal decision rule then takes these predictions as an input and applies group-specific thresholds for admission, which arise to differing social welfare weights on each group G. If the social welfare weight on the disadvantaged group is larger than the social welfare weight on the rest of the population, then the social planner applies a lower threshold for admission for the disadvantaged group.[3]

[3]This intuition is analogous to results in Fryer Jr and Loury (2013), which emphasize the importance

Proposition 1 is quite general. In deriving this result, we placed no assumptions on how the measured outcome $\tilde{Y}$ relates to the outcome of interest Y^*, no assumptions on whether there are group differences in the average outcome of interest conditional on the characteristics $\theta^*(g,w)$ and no assumptions on how the distribution of the characteristics W may differ across groups. Additionally, if the social welfare weights ψ vary not only across groups but across other observable features in W, Proposition 1 generalizes naturally. The first-best algorithm still uses a threshold rule, and the admissions thresholds now vary based upon all characteristics that affect the social welfare weights.

A natural follow-up question is: under what conditions does the social planner use the observed training data D_N in constructing her rank-ordering of the population? Intuitively, we say that the social planner *ignores* the observed training data if her posterior expectation of the outcome of interest equals her prior expectation of the outcome of interest.

Definition 2. *The social planner **ignores** the observed training data D_N 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 with positive probability.*

If the social planner ignores the training dataset, then she learns nothing from the observed training dataset, and therefore, there would be no loss if the social planner discarded it. This holds if and only if the social planner's prior beliefs are such that mismeasured outcome $\tilde{Y}$ is independent of Y^* conditional on the observed characteristics (G,W).

Proposition 2. *The social planner ignores the observed training dataset if and only if under her prior beliefs π, $\tilde{\eta}$ is statistically independent of η.*

This result follows directly from Proposition 1 of Poirier (1998). Intuitively, the likelihood function in Equation (6) only depends on the parameter η through $\tilde{\eta}(g,w)$.[4] Therefore, if the prior beliefs of the social planner are such that the parameters $\tilde{\eta}(g,w)$ are independent of the parameters $\eta^*(g,w)$, then the social planner learns nothing about $\eta^*(g,w)$ from learning about $\tilde{\eta}(g,w)$. This result implies that if the measured outcome $\tilde{Y}$ is statistically related to the outcome of interest Y^* in *any way* under the social planner's beliefs, then the social planner will use the training dataset to construct her optimal algorithm. For example, the social planner may believe the measured outcome $\tilde{Y}$ is mis-measured,

of accurate within-group rankings in designing optimal affirmative action policies (see also, Fryer Jr et al. (2008)).

[4] In other words, the likelihood function in Equation (6) is flat in the parameters $\eta^*(g,w)$ given a particular value of $\tilde{\eta}(g,w)$, and so the parameters of interest are partially identified in this model.

negatively correlated with the outcome of interest, positively correlated with the outcome of interest or "biased" against the disadvantaged group in some way. In all of these cases, the measured outcome $\tilde{Y}$ may still not be independent with the outcome of interest Y^* under the social planner's beliefs, and so it remains optimal to learn from the training dataset.

3.3 Algorithmic decision-making and the first-best admissions rule

An appealing interpretation of Proposition 1 is that the social planner simply constructs an optimal prediction of the measured outcome $\tilde{Y}$ from the observed training data and then uses her prior beliefs π to map these into predictions of the outcome of interest Y^*. This intuition is valid asymptotically as the size of the training dataset grows large.

To develop this result, we first provide a simple definition of a predictive algorithm, which uses the observed training data to construct a prediction function, where the prediction function simply maps observed characteristics (W, G) into predictions of the observed label $\tilde{Y}$.

Definition 3. *A **predictive algorithm** A is a function that maps a training dataset D_N to a prediction function $A(D_N) = \hat{f}_N$, where $\hat{f}_N : \{0,1\}^{J+1} \to [0,1]^K$, where $\hat{f}_{N,k}(g,w)$ is the predicted probability that $\tilde{Y} = \tilde{y}_k$.*

Definition 4. *A predictive algorithm A is **consistent** if its prediction function $\hat{f}_N = A(D_N)$ converges in probability pointwise to the conditional distribution of $\tilde{Y}$ given the characteristics W, G, meaning that as $N \to \infty$*

$$\hat{f}_{N,k}(g,w) \xrightarrow{p} \tilde{\eta}_k(g,w) \quad \forall (g,w) \in \{0,1\}^{J+1} \text{ and } k = 1, \ldots, K.$$

As the size of the training dataset grows large, the social planner's posterior beliefs about $\eta^*(g,w)$ at some fixed characteristics g, w are equivalent asymptotically to the social planner's beliefs if she simply plugged in the predictions of a consistent predictive algorithm to her beliefs about the distribution of the outcome of interest Y^* conditional on the measured outcome $\tilde{Y}$. Define $\pi(\eta^*|\tilde{\eta})$ to be the conditional prior distribution of the parameters η^* given the parameters $\tilde{\eta}$. The social planner's posterior beliefs about η^* are asymptotically equivalent to the beliefs she would have if she plugged in the predictions of a consistent predictive algorithm into her conditional prior beliefs $\pi(\eta^*|\tilde{\eta})$.

Proposition 3. *Let A be a consistent predictive algorithm, and assume that the regularity conditions in Appendix C hold. The social planner's plug-in posterior beliefs $\pi(\eta^*|\hat{f}_N)$ asymptotically*

approximate the social planner's true posterior beliefs $\pi(\eta^*|D_N)$ *as* $N \to \infty$*, meaning*

$$d_{TV}\left(\pi(\eta^*|D_n), \pi(\eta^*|\hat{f}_N)\right) \xrightarrow{p} 0,$$

where $d_{TV}(\cdot,\cdot)$ *denotes the total variation distance between probability measures.*

Proposition 3 implies that the social planner's posterior beliefs are asymptotically equivalent to her beliefs if she constructed a consistent prediction function for the measured outcome in the training dataset and then ex-post mapped these into predictions of the outcome of interest. In other words, to construct her optimal predictive algorithm, the social planner first constructs an accurate predictor for the measured outcome and then modifies them according to her prior beliefs about the relationship between the measured outcome and the outcome of interest.

This result slightly generalizes Theorem 1 in Moon and Schorfheide (2012), which shows that the posterior beliefs of a Bayesian decision-maker about an unidentified parameter given an identified parameter can be approximated asymptotically by their posterior beliefs about the unidentified parameter evaluated at the maximum likelihood estimator for the identified parameter. Proposition 3 shows that the same result holds for any consistent estimator of the identified parameter under the same high-level regularity conditions as Moon and Schorfheide (2012), provided in Appendix C for completeness.

Together, Propositions 1-3 imply a strong-form of *equity irrelevance* - the social planner's equity preferences modify the decision rule but not the predictive algorithm and the only factor in the social planner's choice of predictive algorithm is accuracy. The social planner does not wish to blind the predictive algorithm to group membership, nor remove any characteristics W. Moreover, she does not wish for the predictive algorithm to satisfy any additional fairness constraints that may worsen predictive accuracy. The social planner simply constructs an accurate prediction function of the measured outcome $\tilde{Y}$ using the characteristics W and group membership G. Given this estimated prediction function, the social planner modifies the decision rule in two ways. First, she maps the predictions of the measured outcome into predictions of the outcome of interest using her prior beliefs π and second, she adjusts the admissions thresholds based on the social welfare weights.

3.4 Connections to previous work

Much of the literature in computer science approaches the problem of algorithmic fairness by first introducing a definition of a "fair" prediction function. Given a particular definition, the problem of constructing fair prediction functions reduces to searching for the

most accurate prediction function that satisfies the chosen definition. Because fairness is modelled as an additional constraint in the training procedure, this is commonly referred to as "fairness-constrained" optimization. For example, Dwork et al. (2012) defines a prediction function to be fair if it satisfies a "Lipschitz constraint," which informally means that if two observations have similar observable characteristics, then they should receive similar predictions. Zemel et al. (2013) additionally defines a prediction function to be fair if it satisfies "statistical parity," meaning that the probability that a member of the disadvantaged group is assigned a particular classification is equal to the probability that a member of the non-disadvantaged group is assigned to that same classification.[5] Feldman et al. (2015) formally defines what it means for a prediction function to generate "disparate impact" in terms of classification accuracy across groups and Hardt et al. (2016) introduce two additional notions of fair prediction, which they refer to as "equalized odds" and "equal opportunity." Mitchell et al. (2019) provides a recent review of the wide range of definitions of fairness that exist in the literature.

This approach is crucially different than our analysis of the first-best problem. We did not first introduce a definition of a fair prediction function and then search for the prediction function that maximizes social welfare among all that satisfy the chosen definition. Instead, we began with the social welfare function, which explicitly defines an equity preference in terms of the outcomes of the screening decisions. We placed no restrictions on the admissions rule, and searched among *all* admissions rule to find the optimum. This is a subtle, yet important difference as defining fairness in terms of properties of the underlying prediction function may be unsatisfying for several reasons. First, it is well known that many commonly used definitions of fairness in the computer science literature cannot be simultaneously satisfied (e.g. Raghavan et al., 2017; Chouldechova, 2017; Pleiss et al., 2017). Second, in practice, prediction functions that satisfy a particular definition of predictive fairness may nevertheless produce downstream, unequal outcomes.[6] Given that our preferences for fairness are ultimately defined over downstream outcomes, it is conceptually attractive to directly summarize these preferences as a social welfare function.

Our result in Proposition 1 is most closely related to several recent papers in computer science. Corbett-Davies et al. (2017) show the optimal classifier that satisfies certain

[5]This is sometimes referred to as "group fairness." Kamishima et al. (2011) and Kamishima et al. (2012) introduce regularization techniques that are designed to achieve a similar definition of group fairness.

[6]For example, Liu et al. (2018) highlight that the commonly introduced definitions of fair predictions are static and only describe properties in a single, one-shot prediction exercise. When examined dynamically, the authors show that prediction functions that satisfy, for example, demographic parity may lead to declines in the average predicted outcome for disadvantaged group.

definitions of fairness takes the form of a threshold rule with group-specific thresholds.[7] Lipton et al. (2018) and Menon and Williamson (2018) provide similar results, characterizing the solutions to other "fairness-constrained" loss minimization problems. These are analogous to our result in Proposition 1, except, as mentioned, we show that the same form of the decision rule is *globally* optimal for any social welfare function that takes the form in Equation (2).

Several recent papers in computer science also consider connections between a social welfare approach and existing predictive notions of fairness in computer science. Hu and Chen (2018) consider a related yet different question than the one we pursue. Given a prediction that solves a particular loss minimization problem, the authors characterize the set of social welfare functions that would be optimized by the given prediction function. Similarly, Hu and Chen (2020) assess the welfare impacts of common predictive notions of fairness, where welfare is defined over the resulting outcomes for groups and individuals. Heidari et al. (2018) proposes a training procedure to construct algorithms that minimize some predictive loss subject to a constraint on the average utility of an individual in the population. In contrast, we allow the social planner to explicitly place different weights on payoffs of individuals associated with different groups and assume that the social planner does not value predictive accuracy separately from social welfare. Balashankar et al. (2019) introduce a notion of "pareto-efficient fairness," which searches for prediction functions that jointly maximize predictive accuracy over each group in the population. Building on the welfare framework developed here, Viviano and Bradic (2020) advocate for selecting decision rules that lie on the "pareto frontier" (i.e., decision rules that are not strictly dominated by another in terms of average welfare for any group) and develop statistical techniques for characterizing the fairest decision rule within this frontier. Similarly motivated by our welfare economics framework, Babii et al. (2020) develop computational algorithms to solve classification problems with general loss functions.

Finally, Kleinberg et al. (2018) also introduce an explicit social welfare function that is defined over both the average outcome of admitted individuals and the fraction of admits from the disadvantaged group. Our results differ in two ways. First, the social welfare function in Equation (2) is only defined in terms of the average outcomes of the admitted individuals and not directly on the composition of the admitted class. Second, we explicitly allow for the measured outcome to differ from the outcome of interest.

[7] These fairness definitions are "statistical parity", "conditional statistical parity" and "predictive equality." See Corbett-Davies et al. (2017) for details.

4 Regulating Discrimination and the Detection Problem

In applications in which the social planner selects both the predictive algorithm and the decision rule, our focus on the first-best problem is the relevant policy problem. However, in many other settings, third-party firms or individuals control both the construction of the algorithm and the choice of the admissions rule. Such problems are better modeled as a regulation problem, in which the social planner interacts with a third-party decision-maker and has access to only a limited set of policy instruments to influence their choices. Throughout, we refer to the third-party decision-maker as a *human decision-maker*.

We now extend our model to analyze this regulation problem. The social planner oversees a market of human decision-makers, each of which faces their own screening decision. The human decision-makers have different preferences than the social planner, and some wish to discriminate against the disadvantaged group. The social planner faces a *second-best problem* as she must rely on possibly discriminatory human decision-makers to select admissions decisions that maximize social welfare and may only influence their decisions through policy instruments. Crucially in our analysis of the regulation problem, we focus on modeling the human decision-makers' and social planners' beliefs about what characteristics are predictive of the outcome of interest, putting aside the measured outcome.[8]

Our main results in this section demonstrate that this model captures many of our existing intuitions about regulating discrimination in the absence of algorithms and that the equilibrium level of discrimination in this purely human-driven decision-making environment is strictly positive, highlighting the difficulty of detecting discrimination.

4.1 The market of human decision-makers

There is a market that consists of a unit mass of human decision-makers. Each human decision-maker faces her own screening decision, modeled as in Section 2. A human decision-maker is summarized by three components: preferences $\lambda := (\lambda_0, \lambda_1)$, prior beliefs π_m and a capacity constraint $C \in [0, 1]$.

The human decision-maker's preferences λ govern her payoffs. Similar to the social welfare function, the human decision-maker's payoffs are a weighted average of the outcome of interest among individuals that are admitted into the program

$$U(t; \lambda) := \sum_{(g,w) \in \{0,1\}^{J+1}} \lambda_g \theta^*(g, w) t(g, w) P(g, w), \tag{7}$$

[8]This may be interpreted as assuming that the human decision-makers and the social planner share the same prior beliefs over the relationship between $(\tilde{Y}, Y^*)$ and the social planner simply knows less about what characteristics are predictive of the measured outcome.

where (λ_0, λ_1) are the relative weights placed on the outcomes of each group. If $\lambda_0 > \lambda_1$, then the human decision-maker underweights outcomes associated with the disadvantaged group, leading to the following definition.

Definition 5. *The human decision-maker is* ***discriminatory*** *if* $\lambda_0 > \lambda_1$. *The human decision-maker is* ***non-discriminatory*** *if* $\lambda_0 = \lambda_1$.

Non-discriminatory human decision-makers place equal weight on the outcomes associated with each group, and therefore simply wish to select a decision rule that maximizes the average outcome of interest among the admitted individuals. In this sense, discriminatory human decision-makers are taste-based discriminators in the spirit of Becker (1957). Given the form of the preferences in Equation (7), we normalize $\lambda_0 = 1$ without further loss of generality. We assume there are only two types of preferences in the market: *non-discriminators* with $\lambda = (1, 1)$ and *discriminators* with $\lambda = (1, \bar{\lambda}_1)$ and $\bar{\lambda}_1 < 1$.

Let $m \subseteq \{1, \ldots, J\}$ denote a *model*, where W_m denotes the subvector of $W = (W_1, \ldots, W_J)$ associated with the indices in model m and W_{-m} denotes the subvector of W that is not associated with model m. Let $|m|$ denote the number of characteristics in model m.

The prior beliefs π_m describe the human decision-maker's beliefs about which characteristics $W \in \{0,1\}^J$ are relevant for predicting the outcome of interest Y^* in her screening decision. Each prior π_m is associated with a particular model $m \subseteq \{1, \ldots, J\}$ and is defined such that human decision-makers with prior π_m believe that only the variables in model m contain signal for predicting the outcome of interest Y^*. More concretely, π_m is a joint distribution over the parameters $\{\theta^*(g, w) : (g, w) \in \{0,1\}^{J+1}\}$ satisfying

$$\mathbb{E}_{\pi_m}\left[\theta^*(g, w_m, w_{-m})\right] = \mathbb{E}_{\pi_m}\left[\theta^*(g, w_m, w'_{-m})\right] \tag{8}$$

for all $g \in \{0,1\}, w_m \in \{0,1\}^{|m|}$, $w_{-m}, w'_{-m} \in \{0,1\}^{J-|m|}$. For compactness, write $\theta^*_{\pi_m}(g, w_m) := \mathbb{E}_{\pi_m}\left[\theta^*(g, w)\right]$, where $w = (w_m, w_{-m})$. There are 2^J possible models and there is a prior π_m associated with each model that satisfies Equation (8). The human decision-maker's prior π_m can be thought of as her "mental" algorithm that summarizes which characteristics she believes to be predictive of the outcome of interest.

We assume that at each prior π_m, all characteristics in model m are relevant for predicting the outcome of interest and that the human decision-maker believes that there are no group differences conditional on the characteristics in model m.

Assumption 1 (Sufficiency and relevance)**.** *At each model* $m \subseteq \{1, \ldots, J\}$ *and associated beliefs* π_m*, assume that the characteristics in model* m *are*

(i) ***sufficient****, meaning* $\theta^*_{\pi_m}(0, w_m) = \theta^*_{\pi_m}(1, w_m)$ *for all* $w_m \in \{0,1\}^{|m|}$,

(ii) ***relevant***, *meaning* $\theta^*_{\pi_m}(g, w_m) \neq \theta^*_{\pi_m}(g, w'_m)$ *for all* $w_m, w'_m \in \{0,1\}^{|m|}$ *with* $w_m \neq w'_m$.

With the sufficiency assumption, further write $\theta^*_{\pi_m}(w_m) := \theta^*_{\pi_m}(g, w_m)$, dropping the dependence on group membership.

Finally, each human decision-maker faces a capacity constraint $C \in [0,1]$, meaning the human decision-maker may not admit more than fraction C of the population

$$\sum_{(g,w)\in\{0,1\}^{J+1}} t(g,w)P(g,w) \leq C. \tag{9}$$

The market of human decision-makers is characterized by a joint distribution $\eta(\lambda, \pi_m, C)$ over possible combinations of preferences λ, beliefs π_m and capacity constraints C. This joint distribution has full support, meaning that $\eta(\lambda, \pi_m, C) > 0$ for each possible combination of preferences, beliefs and capacity constraints. We additionally assume that the capacity constraint is independent of preferences and beliefs, meaning that $(\lambda, \pi_m) \perp\!\!\!\perp C$ under η and, therefore we factor this joint distribution into $\eta(\lambda, \pi_m, C) = \eta(\lambda, \pi_m) \times h(C)$.

4.2 The human decision-maker's screening problem

Consider a human decision-maker with preferences λ, beliefs π_m and capacity constraint C. She selects a decision rule that maximizes her expected payoffs subject to the capacity constraint

$$\begin{aligned} \max_{t(g,w)} & \sum_{(g,w)\in\{0,1\}^{J+1}} \lambda_g \theta^*_{\pi_m}(w) t(g,w) P(g,w), \\ \text{s.t.} & \sum_{(g,w)\in\{0,1\}^{J+1}} t(g,w)P(g,w) \leq C. \end{aligned} \tag{10}$$

This is exactly analogous to the social planner's first-best problem in Definition 1. Applying Proposition 1, the human decision-maker's optimal decision rule is a threshold rule that takes the form $\mathbb{1}\left\{\theta^*_{\pi_m}(w) > \tau(g; C, \lambda)\right\}$, in which ties are handled such that the capacity constraint holds with equality. The threshold for admissions $\tau(g; C, \lambda)$ depends on the human decision-maker's preferences. If the human decision-maker is non-discriminatory with $\lambda = (1,1)$, then the threshold is constant across groups. If the human decision-maker is discriminatory with $\lambda = (1, \bar{\lambda}_1)$ and $\bar{\lambda}_1 < 1$, then she applies a higher threshold for admission to the disadvantaged group.

4.3 The social planner's regulation problem

The social welfare function for a given screening problem is defined as before in Equation (2). The social planner's preferences (ψ_0, ψ_1) are *aligned* with the preferences of non-

discriminatory human decision-makers.

Assumption 2 (Alignment)**.** *The social planner's preferences are **aligned** with non-discriminatory human decision-makers at each prior beliefs π_m, meaning that $\psi_0 = \psi_1 = 1$.*

The assumption that the social planner's preferences are aligned with the non-discriminator's preferences is strong. An interpretation is that our model of the regulation problem assumes that a status quo in which the social planner's equity preference is only binding relative to discriminatory human decision-makers and it imposes that the social planner has no additional equity preference relative to the unconstrained choices that would be made by the non-discriminators.

The social planner does not directly observe the preferences λ, the beliefs π_m nor the capacity constraint C of any given human decision-maker. She only knows the joint distribution η of (λ, π_m, C) in the market of human decision-makers. The social planner's payoffs are summarized by the *aggregate* social welfare function

$$\int_C \left(\sum_{(g,w)\in\{0,1\}^{J+1}} \mathbb{E}_\eta \left[\theta^*_{\pi_m}(w) t(g,w) \right] P(g,w) \right) h(C) dC. \tag{11}$$

Given that the social planner's preferences do not equal the discriminators' preferences, the optimal decision rules of human decision-makers will not, in general, maximize the aggregate social welfare function.

4.3.1 Model regulations

The only policy instrument available to the social planner is *model regulations*, meaning that the social planner may regulate what characteristics can be used in decision rules. For example, the social planner may ban the decision rules from explicitly using group membership or it may ban the decision rules from using certain characteristics.[9,10]

Definition 6. *The social planner may place **model regulations** on the human decision-makers' decision rule. If the social planner implements model regulations m, then all decision rules must*

[9]The policy tool of banning certain characteristics from being used by human decision-makers has been considered before in the economics literature on the regulation of insurance markets and pre-existing conditions (Hoy, 1982; Crocker and Snow, 1986; Rothschild, 2011). This policy constraint is consistent with the observation that in practice regulators for example rarely tell firms exactly how many people to hire, that is, where to set admission thresholds.

[10]Our set-up of the regulation problem is similar to the setting studied in Fryer Jr (2009). However, in Fryer Jr (2009), there are no observable characteristics, and therefore, the only policy tools available are quotas on the total level of hiring, whereas here we allow the regulator to directly influence the human decision-makers' hiring rules.

satisfy

$$t(g, w_m, w_{-m}) = t(g, w_m, w'_{-m})$$

for all $g \in \{0,1\}, w_m \in \{0,1\}^{|m|}$ *and* $w_{-m}, w'_{-m} \in \{0,1\}^{J-|m|}$. *If the social planner additionally* ***bans group membership****, then all decision rules must further satisfy, for all* $g, g' \in \{0,1\}$,

$$t(g, w_m, w_{-m}) = t(g', w_m, w'_{-m}).$$

By assuming that the social planner may only place model regulations on human decision-makers, we are restricting the space of policy instruments that is available to the social planner. How the analysis changes under a broader set of potential policy levers is an important topic for future work. Additionally, we are assuming that these model regulations are enforceable and that human decision-makers comply with them in good faith. Assuming that model regulations can be enforced effectively implies that the social planner observes the human decision-makers' decision rules, whereas in practice, the social planner may only observe a finite number of realized admissions decisions. Assuming that human decision-makers comply with model regulations in good faith implies that human decision-makers do not further manipulate characteristics that enter into their chosen model – once a characteristic is used, it is used only in a manner that is consistent with their prior beliefs.

Banning some characteristics from being used in decision rules forces human decision-makers to pool together groups in the population. This may lead human decision-makers to rank-order the population in a way that more closely matches the social planner's preferred rank-ordering. To see this, consider a human decision-maker with preferences λ, beliefs $\pi_{\tilde{m}}$ and capacity constraint C. At model controls m, she now maximizes

$$\sum_{g \in \{0,1\}} \sum_{w_m \in \{0,1\}^{|m|}} \left\{ \lambda_g \mathbb{E}\left[\theta^*_{\pi_{\tilde{m}}}(W_m, W_{-m}) \mid W_m = w_m, G = g \right] \right\} t(g, w_m) P(g, w_m). \tag{12}$$

The human decision-maker rank-orders based upon $\lambda_g \mathbb{E}\left[\theta^*_{\tilde{m}}(W_m, W_{-m}) \mid W_m = w_m, G = g\right]$ as she must pool together individuals that share the same characteristics in model m. Let $t^{\tilde{m}}_{\lambda,C}(g, w; m)$ denote the decision rule that would be selected by a human decision-maker with preferences λ, beliefs $\pi_{\tilde{m}}$ and capacity constraint C at model controls m if she may use group membership.

Similarly, if the social planner additionally bans decision rules from depending on G,

then the human decision-maker maximizes

$$\sum_{w_m \in \{0,1\}^{|m|}} \left\{ \sum_{g \in \{0,1\}} \lambda_g \mathbb{E}\left[\theta^*_{\pi_{\tilde{m}}}(W_m, W_{-m}) \mid W_m = w_m, G = g\right] P(g|w_m) \right\} t(w_m) P(w_m). \tag{13}$$

Since she must further pool individuals across groups, the human decision-maker rank-orders the population using $\sum_g \lambda_g \mathbb{E}\left[\theta^*_{\pi_{\tilde{m}}}(W_m, W_{-m}) \mid W_m = w_m, G = g\right] P(g|w_m)$. Let $t^{\tilde{m}}_{\lambda,C}(w; m)$ denote the decision rule that the human decision-maker would select if she cannot use group membership at model controls m.

The social planner searches over possible model controls to find the one that induces a rank-ordering most closely aligned with her first-best rank-ordering. This is the *second-best problem*.

Definition 7. *The social planner's **second-best problem** is to select the model regulations that maximize aggregate social welfare, taking the decision rules chosen by human decision-makers as given. She solves*

$$m^* = \arg \max_{m \subseteq \{1,\ldots,J\}} \int_C \left(\sum_{(g,w) \in \{0,1\}^{J+1}} \mathbb{E}_\eta \left[\theta^*_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{\lambda,C}(g, w; m) P(g, w)\right] \right) h(C) dC.$$

If she additionally bans human decision-makers from using group membership, she solves

$$m^* = \arg \max_{m \subseteq \{1,\ldots,J\}} \int_C \left(\sum_{(g,w) \in \{0,1\}^{J+1}} \mathbb{E}_\eta \left[\theta^*_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{\lambda,C}(w; m) P(g, w)\right] \right) h(C) dC.$$

*The solution m^* is the social planner's **second-best model regulations**.*

Finally, the "level of discrimination" at model controls m equals the fraction of discriminatory human decision-makers that select a decision rule that is different than the decision-rule chosen by non-discriminatory human decision-makers with the same beliefs and capacity constraint.

Definition 8. *The **level of discrimination** at model controls m equals*

$$\Delta(m) := \mathbb{P}\left\{ t^{\tilde{m}}_{\lambda,C}(m) \neq t^{\tilde{m}}_{(1,1),C}(m) \mid \lambda = (1, \bar{\lambda}_1) \right\},$$

*where $\mathbb{P}\left\{\cdot \mid \lambda = (1, \bar{\lambda}_1)\right\}$ is the conditional joint distribution of beliefs $\pi_{\tilde{m}}$ and the capacity constraint C among discriminatory human decision-makers. The **equilibrium level of discrimination** is $\Delta(m^*)$.*

4.4 Characterizing the social planner's second-best model regulations

We now characterize the social planner's second-best model regulations when she is faced with human decision-makers. To do so, we formalize what it means for the group $G = 1$ to be disadvantaged. Disadvantage in this model means that characteristics associated with lower average values of the outcome of interest are more likely to occur among the disadvantaged group.

Assumption 3 (Disadvantage condition)**.** *At each beliefs π_m, if w, w' are such that $\theta^*_{\pi_m}(w) \geq \theta^*_{\pi_m}(w')$, then*

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

and this holds with strict inequality if $\theta^*_{\pi_m}(w) > \theta^*_{\pi_m}(w')$.

Together, the disadvantage condition (Assumption 3) and the sufficiency condition (Assumption 1) imply that, conditional on all features at a given model, there are no average differences in the outcome between members of the advantaged and disadvantaged group yet features associated with lower average levels of the outcome of interest are more likely to be observed among the disadvantaged group. Put in another way, we are assuming that disadvantage in the model only arises through the distribution of features across groups.

How exactly the social planner selects model regulations may be quite complex. It will depend on the relative fractions of discriminatory and non-discriminatory human decision-makers as well as the distribution of beliefs π_m across the market of human decision-makers. Therefore, in order to build intuition, we start by considering two simpler problems.

First, suppose that there are only non-discriminatory human decision-makers in the market and that all human decision-makers have the same beliefs $\pi_{\tilde{m}}$. In this case, provided the disadvantage condition is satisfied at model $\tilde{m}$, the social planner lets the human decision-makers use any model m that satisfies $\tilde{m} \subseteq m$, meaning that it includes all characteristics that are believed to be predictive of the outcome of interest.

Proposition 4. *Suppose that there are only non-discriminatory human decision-makers with model $\tilde{m}$ in the market. Then, the social planner's second-best regulation m^* may be any model in the set $\{m : \tilde{m} \subseteq m\}$ and the social planner is indifferent to banning group membership G.*

Under Assumption 2, the social planner's preferences are sufficiently aligned with the non-discriminatory human decision-maker's preferences such that the social planner does not wish to change the rank-ordering of the non-discriminatory human decision-maker.

Banning characteristics that are believed to be predictive of the outcome of interest only introduces mis-rankings that lower aggregate social welfare.

Next, suppose that there are only discriminatory human decision-makers in the market and that all human decision-makers have the same beliefs $\pi_{\tilde{m}}$. Intuitively, discriminatory human decision-makers have sufficiently different preferences than the social planner that the social planner may find it optimal to place model regulations. Indeed, provided that the disadvantage condition is satisfied at beliefs $\pi_{\tilde{m}}$, this is true – it is optimal for the social planner to implement model controls, forcing the discriminatory human decision-makers to use model $\tilde{m}$ and ban them from using group membership.

Proposition 5. *Suppose that there are only discriminatory human decision-makers with model $\tilde{m}$ in the market and the disadvantage condition holds. Then, it is optimal for social planner to place model controls that force the human decision-makers to use model $\tilde{m}$ and ban group membership.*

The proof shows that at these model controls, the rank-ordering used by discriminatory human decision-makers is the same as the rank-ordering used by non-discriminatory human decision-makers. This result is reminiscent of a "disparate treatment" test because the social planner wishes to force discriminatory human decision-makers to treat members of both groups the same given the characteristics in model $\tilde{m}$.

To this point, we considered special cases in which all human decision-makers had the same beliefs and the regulator knew those beliefs exactly. In general, there is an entire market of human decision-makers with different beliefs about which characteristics are predictive of the outcome of interest. This additional dimension of private information induces a trade-off. Banning group membership creates an incentive for a discriminatory human decision-maker to use more characteristics in her decision rule than she believes to be predictive of the outcome of interest in order to screen out members of the disadvantaged group. In other words, it creates incentives for discriminatory human decision-makers to select decision rules that generate *disparate impact*.

Proposition 6. *Consider a discriminatory human decision-maker with model $\tilde{m}$ and assume that $P(G = 1|W = w) \neq P(G = 1|W = w')$ for all $w, w' \in \{0,1\}^J$ with $w \neq w'$. If group membership G is banned, then the discriminatory human decision-maker's optimal decision rule is based on a rank-ordering that uses all characteristics $W \in \{0,1\}^J$.*

Provided that the social planner bans group membership from being used in decision rules, human decision-makers may wish to use an additional characteristic for two reasons. Some human decision-makers may believe that it is predictive of the outcome of interest and others may wish to use it in order to screen out the disadvantaged group.

This intuition produces the *flexibility tradeoff* in regulating discrimination. Letting human decision-makers use more characteristics leads to more accurate rank-orderings of the population but it also makes it easier for discriminatory human decision-makers to screen out the disadvantaged group.

Proposition 7. *Suppose that the social planner bans human decision-makers from using group membership in their decision rules. Then, the second-best problem in Definition 7 is equivalent to*

$$\min_{m\subseteq\{1,\ldots,J\}} \sum_{\tilde{m}} \left\{ \int_C \mathbb{E}\left[\psi(W)\theta^*_{\pi_{\tilde{m}}}(W)\left(t^{\tilde{m}}_{*,C}(W) - t^{\tilde{m}}_{ND,C}(W;m)\right)\right] h(C)dC \right\} \eta(\tilde{m}) \\ + \eta(D) \sum_{\tilde{m}} \left\{ \int_C \mathbb{E}\left[\psi(W)\theta^*_{\pi_{\tilde{m}}}(W)\left(t^{\tilde{m}}_{ND,C}(W;m) - t^{\tilde{m}}_{D,C}(W;m)\right)\right] h(C)dC \right\} \eta(\tilde{m}|D),$$

where $\psi(W) = \psi_0 P(0|w) + \psi_1 P(1|w)$ *and* $t^{\tilde{m}}_{*,C}(W)$ *denotes the social planner's first-best decision rule at beliefs* $\pi_{\tilde{m}}$ *and capacity constraint* C*.*

The first term in Proposition 7 depends on the difference between the social planner's first-best decision rule at beliefs $\pi_{\tilde{m}}$ and the decision rule that non-discriminatory human decision-makers with beliefs $\pi_{\tilde{m}}$ would select at model controls m. Under the alignment assumption (Assumption 2), for model controls satisfying $\tilde{m} \subseteq m$, the rank-order used by the the non-discriminatory human decision-maker matches the social planner's first-best rank-order. Therefore, as the number of characteristics allowed grows, the first term declines to zero, capturing the gains from more accurate rank-ordering. The second term depends on the difference between the decision rule selected by non-discriminatory human decision-makers and discriminatory human decision-makers at the same beliefs $\pi_{\tilde{m}}$ and model controls m. These differ only because of the different preferences λ between these human decision-makers. Once the model controls are such that $\tilde{m} \subset m$, the decision rule selected by the non-discriminatory human decision-makers no longer changes but the discriminatory human decision-makers now use any extra features to screen out members of the disadvantaged group (Proposition 6). As the number of allowed characteristics increases, there are more differences between the decision rules of the non-discriminatory human decision-makers and discriminatory human-decision-makers, lowering social welfare. This effect captures the intuition that more permissive model regulations makes it easier for discriminatory human decision-makers to select decision rules that generate disparate impact.

Finally, we show that the equilibrium level of discrimination is non-zero in the second-best problem provided that there is a conflict in the preferred ranking-orderings of discriminatory and non-discriminatory human decision-makers.

Proposition 8. *Suppose that for all beliefs $\pi_{\tilde{m}}$, there exists a pair of characteristics $w_{\tilde{m}}, w'_{\tilde{m}}$ such that*

$$\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}), \text{ and } \bar{\lambda}(w')\theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}) > \bar{\lambda}(w)\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}),$$

where $\bar{\lambda}(w_{\tilde{m}}) = P(0|w_{\tilde{m}}) + \bar{\lambda}_1 P(1|w_{\tilde{m}})$. Then, the equilibrium level of discrimination is strictly positive with $\Delta(m^*) > 0$.

Because the social planner must select a single model regulation for the entire market, there always exists some discriminatory human decision-makers that are given sufficient freedom to select a decision rule that differs from the corresponding non-discriminatory human decision-maker. In equilibrium, discrimination goes undetected. The stated condition in Proposition 8 imposes that discriminatory preferences induce a wedge in the preferred ranking-orderings.

5 Algorithmic Decision-Making and Second-Best Model Regulations

To this point, we considered the social planner's second-best problem when she oversees a market of human decision-makers without algorithms. The social planner faced two sources of asymmetric information: over the preferences λ and over the beliefs π_m of the human decision-makers and *both* dimensions of asymmetric information gave rise to the flexibility tradeoff.

We now consider what happens when human decision-makers adopt predictive algorithms in their screening decisions. How this affects the social planner's second-best model regulations depends crucially on what human decision-makers must disclose about their predictive algorithms and decisions rules. First, we consider a full disclosure regime in which human decision-makers' are subject to algorithmic audits, meaning that they must disclose both their decision rule and predictive algorithm to the social planner. In this case, the social planner now finds it optimal to let any characteristic that is predictive of the outcome of interest be used in decision rules and the equilibrium level of discrimination is zero. Second, to highlight the importance of full disclosure, we consider the case in which human decision-makers' only disclose their decision rule but not their predictive algorithm. In this case, optimal regulation is the same as the case with a purely human-driven decision loop.

5.1 Introducing algorithmic decision-making

We model the introduction of predictive algorithms as revealing the ground truth $\theta^*(g, w)$ in each screening problem to the human decision-makers. An interpretation is that the

human decision-makers receive access to a large, randomly sampled dataset from the population of individuals and using this training dataset to train a consistent predictive algorithm (Definition 4). Formally, each human decision-maker is now associated with a *ground-truth model*.

Definition 9. *A **ground-truth model** m summarizes the set of characteristics that are relevant in predicting the outcome of interest in a screening problem. It is associated with parameters $\mathbb{E}[Y^* \mid G = g, W = w] = \theta^*(g, w)$ that satisfy*

$$\theta^*(g, w_m, w_{-m}) = \theta^*(g', w_m, w'_{-m})$$

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

The ground-truth model m is the human decision-maker's predictive algorithm. At the ground-truth model m, the characteristics W_m are sufficient and relevant for predicting the outcome of interest in the population of individuals. Denote the average outcome of interest at ground-truth model m as $\theta^*(g, w_m, w_{-m}) := \theta^*_m(w_m)$ for all $g \in \{0,1\}, w_{-m} \in \{0,1\}^{J-|m|}$.

Given their ground-truth model, preferences and capacity constraint, each human decision-maker selects a decision rule to maximize their payoffs, which are now defined as

$$\sum_{(g,w)\in\{0,1\}^{J+1}} \lambda_g \theta^*_m(w) t(g,w) P(g,w). \tag{14}$$

The human decision-maker's optimal decision rule is a threshold rule $\mathbb{1}\{\theta^*_m(w) > \tau(g; C, \lambda)\}$, in which ties are handled such that the capacity constraint holds with equality and the threshold $\tau(g; C, \lambda)$ may vary across groups.

Finally, the market of human decision-makers is now summarized by a joint distribution over ground-truth models m, preferences λ and capacity constraints C and in a slight abuse of notation, we again denote this joint distribution by η. We continue to assume that the distribution has full support and that the capacity constraint is independent of the ground-truth model and preferences in the market of human decision-makers, meaning $(m, \lambda) \perp\!\!\!\perp C$ under η.

5.2 Second-best model regulations in the presence of algorithmic audits

The adoption of predictive algorithms introduces a new policy tool to the social planner – *algorithmic audits*. An algorithmic audit refers to the process in which the social planner

may access the underlying training data and training procedure that the human decision-maker used to construct her algorithm. Kleinberg et al. (2018, 2020) describe in detail how algorithmic audits may function in practice. We model algorithmic audits in a reduced-form manner as simply revealing the ground-truth model m of each human decision-maker to the social planner.

Definition 10. *An **algorithmic audit** reveals the ground-truth model θ^*_m of each human decision-maker in the market.*

If the social planner may implement algorithmic audits, then the adoption of predictive algorithms eliminates one dimension of private information between the social planner and the human decision-makers. She may now condition her model regulations on the ground-truth model m. This has important ramifications for how the social planner sets her optimal model regulations.

In the presence of algorithmic audits, the social planner's second-best problem is now to select her model regulations that maximize aggregate social welfare, conditional on the ground-truth model m revealed by the algorithmic audit.

Definition 11. *Suppose the social planner may conduct algorithmic audits. The social planner's **algorithmic second-best problem** is to select model regulations that maximize aggregate social welfare among all human decision-makers with ground-truth model m, taking the decision rules chosen by the human decision-makers as given. That is, she solves*

$$m^*(m) = \arg \max_{\tilde{m} \subseteq \{1,\dots,J\}} \int_C \left(\sum_{(g,w) \in \{0,1\}^{J+1}} \mathbb{E}_{\lambda|m} \left[\theta^*_m(w) t^m_{\lambda,C}(g,w;\tilde{m}) P(g,w) \right] \right) h(C) dC,$$

where $\mathbb{E}_{\lambda|m}[\cdot]$ is an expectation over conditional distribution of preferences given the true model m, $\eta(\lambda|m)$.

Our earlier results from Section 4 immediately imply that the social planner's second-best algorithmic regulations are simple if she may conduct algorithmic audits. At ground-truth model m, the social planner finds it optimal to set model controls $\tilde{m} = m$ and ban group membership. We state this in the next proposition.

Proposition 9. *In the presence of algorithmic audits, a second-best model regulation for the social planner allows the human decision-makers decision-makers to use any characteristics that is predictive of the outcome of interest and bans group membership. That is, $m^*(\tilde{m}) = \tilde{m}$ for all ground-truth models $\tilde{m}$.*

Proof. This result follows immediately from Proposition 4 and Proposition 5, which imply that an optimum for the social planner is to set $m^*(\tilde{m}) = \tilde{m}$ and ban group membership G from being used in decision rules. □

The intuition underlying this result is quite simple. If there were only non-discriminators among human decision-makers with ground-truth model $\tilde{m}$, then the social planner would find it optimal to select any model controls m satisfying $\tilde{m} \subseteq m$. If there were only discriminators among the human decision-makers with ground-truth model $\tilde{m}$, then the social planner would find it optimal to select model controls $m = \tilde{m}$ and ban the use of group membership. Proposition 9 follows immediately from these two results.

Moreover, the presence of algorithmic audits has strong implications for the equilibrium level of discrimination. If the social planner may conduct algorithmic audits, then the introduction of algorithmic decision-making *lowers* the equilibrium level of discrimination relative to its level without algorithms and in fact, the equilibrium level of discrimination goes to zero provided that the disadvantage condition holds.

Proposition 10. *If the social planner may conduct algorithmic audits, then the equilibrium level of discrimination is zero (i.e., $\Delta(m^*) = 0$).*

Because the social planner no longer faces asymmetric information over both the human decision-makers' preferences and the ground truth model if she can conduct algorithmic audits, she is able to force discriminatory human decision-makers to select the same decision rule as non-discriminatory human decision-makers. This highlights a core gain from the adoption of predictive algorithms – there is a reduction in the level of discrimination provided that the social planner may conduct algorithmic audits.

5.3 Second-best model regulations with known decision rules

Finally, we consider a disclosure regime in which human decision-makers must only disclose their decision rule to the social planner. In this case, the introduction of predictive algorithms does not change the social planner's second-best regulation problem. Since she still faces asymmetric information over the ground-truth model of the human decision-makers, the social planner still faces the flexibility tradeoff in Proposition 7, highlighting the importance of full disclosure of both the ground-truth model and the decision rule in the previous regime. If human decision-makers must only disclose their decision rule, optimal regulation does not change relative to the case with a purely human-driven decision loop.

Proposition 11. *Suppose that the human decision-makers adopt algorithms and the social planner bans human decision-makers from using group membership in their decision rules. Then, the social*

planner's second-best problem is again equivalent to

$$\min_{m\subseteq\{1,\ldots,J\}} \sum_{\tilde{m}} \left\{ \int_C \mathbb{E}\left[\theta^*_{\tilde{m}}(W)\left(t^{\tilde{m}}_{*,C}(W) - t^{\tilde{m}}_{ND,C}(W;m)\right)\right] h(C)dC \right\} \eta(\tilde{m}) \\ + \eta(D) \sum_{\tilde{m}} \left\{ \int_C \mathbb{E}\left[\theta^*_{\tilde{m}}(W)\left(t^{\tilde{m}}_{ND,C}(W;m) - t^{\tilde{m}}_{D,C}(W;m)\right)\right] h(C)dC \right\} \eta(\tilde{m}|D),$$

where $t^{\tilde{m}}_{*,C}(W)$ denotes the social planner's first-best decision rule at ground-truth model $\tilde{m}$ and capacity constraint C.

Corollary 1. *If at each ground truth model $\tilde{m}$, there exists a pair of characteristics $w_{\tilde{m}}, w'_{\tilde{m}}$ such that*

$$\theta^*_{\tilde{m}}(w_{\tilde{m}}) > \theta^*_{\tilde{m}}(w'_{\tilde{m}}), \text{ and } \bar{\lambda}(w')\theta^*_{\tilde{m}}(w'_{\tilde{m}}) > \bar{\lambda}(w)\theta^*_{\tilde{m}}(w_{\tilde{m}}),$$

then the equilibrium level of discrimination is strictly positive with $\Delta(m^*) > 0$.

Since the social planner can still only observe the decision rule selected by the human decision-maker, she is still unsure of *why* this decision rule was selected. Non-discriminatory human decision-makers may be using a characteristic in their decision rule because it is predictive of the outcome of interest at their ground-truth model. In contrast, discriminatory human decision-makers may be using a characteristic in their decision rule because it helps screen out members of the disadvantaged group. In this disclosure regime, this asymmetric information problem is still present. Moreover, as before, the equilibrium level of discrimination is positive if the discriminatory preferences are binding at each ground truth model.

6 Conclusion

We developed an economic model of screening decisions that embeds concerns about algorithmic bias within a social welfare function. The social welfare function depends directly on the outcomes of the screening decision, in which individuals from a population are screened into a program based on predictions of an unknown outcome of interest.

We first considered the social planner's first-best problem, in which the social planner constructed a prediction function and selected the decision rule. The social planner's first-best decision rule ranks the population using all available information and then admits individuals according to that ranking with group-specific admissions thresholds. The social planner's posterior beliefs are asymptotically equivalent to her beliefs if she constructed a consistent predictor of the measured outcome in the training dataset and ex-post mapped these into predictions of the outcome of interest. These results highlight

a strong form of equity irrelevance – equity preferences only modify the decision rule, not the prediction function, in the first-best problem.

Next, we considered the social planner's second-best problem, in which the social planner regulates the screening decisions of human decision-makers with possibly different preferences. The social planner faces a flexibility tradeoff – allowing human decision-makers to use more characteristics leads to more accurate predictions but it also enables discriminatory human decision-makers to screen out the disadvantaged group. Discrimination goes undetected, as the equilibrium level of discrimination is strictly positive. With algorithmic decision-making, the social planner may learn the true prediction function used by human decision-makers through algorithmic audits. In this case, the social planner lets the human decision-makers use any characteristic that contains signal in predicting the outcome of interest. Moreover, with algorithmic audits in place, the equilibrium level of discrimination declines, highlighting that algorithmic decision-making not only improves prediction but may also make it easier to detect discrimination.

Our results analyzed the optimal use and regulation of algorithmic decision rules under a specific set of assumptions about the social planner, private actors and the nature of their interaction. It would be useful to enrich our analysis by allowing for more forms of discriminatory behavior and richer forms of information asymmetries between the social planner and the firms. For example, our analysis considered the case in which some human decision-makers were tasted-based discriminators, ruling out other possible forms of discriminatory behavior (Fang and Moro, 2011; Bordalo et al., 2016). We also assumed that the social planner and firms agreed on the outcome of interest and shared a common prior. Finally, we abstracted away from finite-sample issues in algorithmic audits, assuming that the human decision-makers and social planner learned ground truth once they accessed an algorithm. Analyzing the second-best problem in full generality is an important task moving forward, requiring insights from both economics and computer science.

References

Arnold, D., W. Dobbie, and P. Hull (2020a). Measuring racial discrimination in algorithms. Technical report, NBER Working Paper No. 28222.

Arnold, D., W. Dobbie, and P. Hull (2020b). Measuring racial discrimination in bail decisions. Technical report, NBER Working Paper No. 26999.

Arnold, D., W. Dobbie, and C. Yang (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics 133*(4), 1885—-1932.

Athey, S. C., K. A. Bryan, and J. S. Gans (2020). The allocation of decision authority to human and artificial intelligence. Technical report, NBER Working Paper No. 26673.

Babii, A., X. Chen, E. Ghysels, and R. Kumar (2020). Binary choice with asymmetric loss in a data-rich environment: Theory and an application to racial justice. Technical report, arXiv preprint, arXiv:2010.08463.

Balashankar, A., A. Lees, C. Welty, and L. Subramanian (2019). What is fair? exploring pareto-efficiency for fairness constrained classifiers. Technical report, arXiv preprint arXiv:1910.14120.

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Becker, G. (1957). *The Economics of Discrimination*. University of Chicago Press.

Blum, A. and K. Stangl (2019). Recovering from biased data: Can fairness constraints improve accuracy? *CoRR abs/1912.01094*.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data 5*(2).

Chouldechova, A. and A. Roth (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM 63*(5), 82–89.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd Conference on Knowledge Discovery and Data Mining*.

Cowgill, B. and M. Stevenson (2020). Algorithmic social engineering. Technical report.

Cowgill, B. and C. Tucker (2019). Economics and algorithmic fairness. Technical report.

Crocker, K. J. and A. Snow (1986). The efficiency effects of categorical discrimination in the insurance industry. *The Journal of Political Economy 94*(2), 321–344.

Doleac, J. and M. Stevenson (2019). Algorithmic risk assessment in the hands of humans. Technical report, IZA Discussion Paper Series No. 12853.

Dwork, C., T. P. Moritz Hardt, O. Reingold, and R. Zemel (2012). Fairness through awareness. *ITCS 2012 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Fang, H. and A. Moro (2011). Theories of statistical discrimination and affirmative action: A survey. In J. Benhabib, M. O. Jackson, and A. Bisin (Eds.), *Handbook of Social Economics*, Volume 1A, pp. 133–200. North-Holland.

Feldman, M., S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Fryer Jr, R. (2009). Implicit quotas. *The Journal of Legal Studies 38*(1), 1–20.

Fryer Jr, R. and G. Loury (2013). Valuing diversity. *Journal of Political Economy 121*(4), 747–774.

Fryer Jr, R., G. Loury, and T. Yuret (2008). An economic analysis of color-blind affirmative action. *Journal of Law, Economics, and Organization 24*(2), 319–355.

Hardt, M., E. Price, and N. Srebro (2016). Equality of opportunity in supervised learning. *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

Heidari, H., C. Ferrari, K. P. Gummadi, and A. Krause (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 1273–1283.

Hoy, M. (1982). Categorizing risks in the insurance industry. *The Quarterly Journal of Economics 97*, 321–336.

Hu, L. and Y. Chen (2018). Welfare and distributional impacts of fair classification. *FAT/ML workshop at the 35th International Conference on Machine Learning*.

Hu, L. and Y. Chen (2020). Fair classification and social welfare. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 535–545.

Kallus, N. and A. Zhou (2018). Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*.

Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma (2012). Fairness-aware classifier with prejudice remover regularizer. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Part II*, 35–50.

Kamishima, T., S. Akaho, and J. Sakuma (2011). Fairness-aware learning through regularization approach. *2011 IEEE 11th International Conference on Data Mining Workshops*.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review: Papers and Proceedings 105*(5), 491–495.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018). Algorithmic fairness. *AEA Papers and Proceedings 108*, 22–27.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. Sunstein (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis 80*, 1–62.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. R. Sunstein (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*.

Lipton, Z., J. McAuley, and A. Chouldechova (2018). Does mitigating ml's impact disparity require treatment disparity? *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Liu, L. T., S. Dean, E. Rolf, M. Simchowitz, and M. Hardt (2018). Delayed impact of fair machine learning. *Proceedings of the 35 th International Conference on Machine Learning*.

Menon, A. K. and R. C. Williamson (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2019). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Technical report, arXiv Working Paper, arXiv:1811.07867.

Moon, H. R. and F. Schorfheide (2012). Bayesian and frequentist inference in partially identified models. *Econometrica 80*(2), 755–782.

Mullainathan, S. and Z. Obermeyer (2017). Does machine learning automate moral hazard and error? *American Economic Review: Papers & Proceedings 107*(5), 476—-480.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453.

Passi, S. and S. Barocas (2019). Problem formulation and fairness. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger (2017). On fairness and calibration. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory 14*(4), 483–509.

Raghavan, M., J. Kleinberg, and S. Mullainathan (2017). Inherent trade-offs in the fair determination of risk scores. *The 8th Innovations in Theoretical Computer Science Conference*.

Rambachan, A., J. Kleinberg, J. Ludwig, and S. Mullainathan (2020). An economic perspective on algorithmic fairness. *AEA Papers and Proceedings 110*, 91–95.

Rambachan, A. and J. Roth (2020). Bias In, Bias Out? Evaluating the Folk Wisdom. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, Volume 156, pp. 6:1–6:15.

Rothschild, C. (2011). The efficiency of categorical discrimination in insurance markets. *The Journal of Risk and Insurance 78*(2), 267–285.

Viviano, D. and J. Bradic (2020). Fair policy targeting. Technical report, arXiv preprint, arXiv:2005.12395.

Wang, H., H. Hsu, M. Diaz, and F. P. Calmon (2020). To split or not to split: The impact of disparate treatment in classification. Technical report, arXiv preprint, arXiv:2002.04788.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning 28*(3), 325–333.

An Economic Approach to Regulating Algorithms

Online Appendix

Ashesh Rambachan Jon Kleinberg Sendhil Mullainathan Jens Ludwig

In the online appendix, we collect together some additional results and assumptions that are discussed in the main text. We also provide the proofs of the main results.

A Common sources of bias

In this section, we discuss how the first-best algorithm design problem in Sections 2-3 and the second-best regulation problem in Section 4 relate to commonly discussed sources of algorithm bias. We summarize the discussion in Table 1 below, which highlights the key assumptions we make about the underlying data generating process in our analysis of the first-best algorithm design problem and the regulation problem. We also discuss several points directly.

The outcome of interest is mis-measured: An increasingly important concern in the literature on algorithmic decision-making and fairness centers on possible mis-measurement in the outcome of interest. For example, Obermeyer et al. (2019) analyze an algorithmic decision tool that generated large racial disparities across patients, which arose because the underlying prediction function predicted historical cost of care as opposed to a measure of the patient's health. See Mullainathan and Obermeyer (2017); Passi and Barocas (2019); Kleinberg et al. (2020); Rambachan et al. (2020) for further discussion of mis-measured outcomes. Both our analysis of the first-best algorithm design problem and the second-best regulation problem allows for a difference between the measured outcome $\tilde{Y}$ and the outcome of interest Y^*. In the first-best problem, the social planner specifies her prior beliefs over the relationship between $(\tilde{Y}, Y^*)$ and uses those beliefs when making her screening decisions. In the regulation problem, both the social planner and firms now specify prior beliefs over $(\tilde{Y}, Y^*)$, and we further require that both the social planner and the firms share common beliefs about the joint distribution of $(\tilde{Y}, Y^*)$.

Characteristics W are correlated with group membership: A key concern in existing research on algorithmic fairness focuses on whether the characteristics W are differently distributed across groups. This could arise because the observable characteristics are measured differently across groups or are themselves the result of prior discrimination. For example, in credit scoring, prior credit history and past income may be correlated with group membership due to discrimination in the credit and labor market. An extreme version occurs when group membership may be perfectly reconstructed using the characteristics W, which is termed the "reconstruction problem" in Kleinberg et al. (2018) or algorithmic "redlining" in Dwork et al. (2012).

Our analyses of the first-best algorithm design problem and the regulation problem allows for the characteristics W to be distributed differently across groups. This allows for the characteristics W that are associated with higher values of the outcome of interest Y^* to be more common among the group $G = 1$ than the group $G = 0$ (for example, the disadvantaged condition in Assumption 3). However, we rule out that the characteristics W may perfectly reconstruct group memberships G. This is implied by our assumption that $P(g, w) > 0$ for all $g \in \{0, 1\}$ and $w \in \{0, 1\}^J$, which means $0 < P(g|w) < 1$ for all $g \in \{0, 1\}$ and $w \in \{0, 1\}^J$. This restriction is important in our analysis of the regulation problem. If group membership can be perfectly reconstructed from observable characteristics, then discriminatory firms can implement their preferred discriminatory decision-rule through a decision-rule that is superficially group-blind. This would imply that the social planner would wish to regulate the use of characteristics W, in addition to merely regulating group membership, thereby breaking our stepping stone result in Proposition 5.

The protected group is "under-represented" in the training data: A common concern is that the protected group may be "under-represented" in the training data, meaning that there may be substantially fewer observations associated with the protected group than the rest of the population. This could arise for several reasons.

First, if the protected group is a minority group in the population of interest, then training data that is generated by taking a random sample from the population of interest would mechanically "under-represent" the protected group with high probability. This may be important since it means that predictions for the protected group may be noisier than the rest of the population (e.g., see Wang et al. (2020) and Blum and Stangl (2019)). Our analysis of the first-best algorithm design problem allows for this form of under-representation in the training data, and illustrates that this does not pose a challenge for the social planner. Social welfare is maximized on average provided the social planner makes admissions decisions according to her best estimate of the rank-ordering of the population in terms of $E[Y \mid W, G]$. In finite sample, the social planner takes the training data as given, updates to her posterior beliefs and admits individuals according to her posterior beliefs according to $E[Y \mid W, G]$ (Proposition 1). As the size of the training data grows large, the social planner can do no better than constructing a consistent estimator of $E[Y \mid W, G]$ (Proposition 3).

Second, the protected group may be under-represented in the training data because the training data may have been generated through a discriminatory selection process. For example, in pre-trial release decisions, bail judges may discriminate by releasing more white defendants than minority defendants (Arnold et al., 2018, 2020b). Since we may only observe whether a defendant commits pre-trial misconduct if they were released, the judge's discrimination means that we observe fewer minority defendants in the training data. Analogous discriminatory selection arises in hiring decisions and lending decisions as well. Since we assume throughout the paper that the observable training data is a random sample from the population of interest, we are ruling out this form of under-representation. Exploring how possibly discriminatory selection affects the evaluation and design of algorithmic decision rules is an active area of research – see, for example,

Kallus and Zhou (2018), Rambachan and Roth (2020) and Arnold et al. (2020a).

Source of Bias	First-Best Algorithm Design	Regulation Problem
Outcomes		
Measured $\tilde{Y}$ differs from outcome of interest Y^*	✓	✓
Group G has direct effect on outcomes $\tilde{Y}, Y^*$	✓	×
Characteristics		
Characteristics W correlated with group G	✓	✓
Characteristics W perfectly "reconstruct" group	×	×

Table 1: This table summarizes how several commonly discussed sources of algorithmic bias are related to the analysis of the first-best algorithm design problem in Sections 2-3 and the regulation problem in Section 4.

B Motivating the social welfare function

In this section, we sketch a brief motivation for the social welfare function given in Equation 2.

As in the main text, let $\tilde{Y} \in \{\tilde{y}_1, \ldots, \tilde{y}_K\}$ denote the measured outcome and let $u_g(\tilde{Y}; T)$ denote the utility of an individual in group g with measured outcome $\tilde{Y}$ that is assigned to the program $T \in \{0, 1\}$. Write this as

$$u_g(\tilde{Y}; T) = T \cdot u_g(\tilde{Y}; 1) + (1 - T) \cdot u_g(\tilde{Y}; 0).$$

Therefore, at decision rule $t(g, w) \in [0, 1]$, an individual's expected utility at a fixed measured outcome $\tilde{Y}$ is

$$t(g, w) \cdot u_g(\tilde{Y}; 1) + (1 - t(g, w)) \cdot u_g(\tilde{Y}; 0),$$

where $t(g, w)$ is the probability that an individual with characteristics (g, w) is assigned to the program.

The social welfare function is a weighted average of individual expected utilities under the decision rule

$$\sum_{(g,w) \in \{0,1\}^{J+1}} \psi_g \left\{ \sum_{k=1}^{K} \left(t(g, w) \cdot u_g(\tilde{y}_k; 1) + (1 - t(g, w)) \cdot u_g(\tilde{y}_k; 0) \right) P(\tilde{y}_k | g, w) \right\} P(g, w),$$

where $P(\tilde{y}_k | g, w) = \mathbb{P}\left\{ \tilde{Y} = \tilde{y}_k \,|\, G = g, W = w \right\}$ and (ψ_0, ψ_1) are generalized social welfare weights that vary across groups. Defining $\Delta_g(\tilde{Y}) := u_g(\tilde{Y}; 1) - u_g(\tilde{Y}; 0)$, it is imme-

diate that maximizing social welfare is equivalent to maximizing

$$\sum_{(g,w)\in\{0,1\}^{J+1}} \psi_g \left\{ \sum_{k=1}^{K} \Delta_g(\tilde{y}_k) P(\tilde{y}_k|g,w) \right\} t(g,w)P(g,w) =$$
$$\sum_{(g,w)\in\{0,1\}^{J+1}} \psi_g \mathbb{E}\left[\Delta_g(\tilde{Y}) \mid G=g, W=w\right] t(g,w)P(g,w)$$

Therefore, without loss of generality, we may redefine the social welfare function as this object. Setting the outcome of interest to be $Y^* = \Delta_g(\tilde{Y})$ delivers the social welfare function given in Equation 2. The outcome of interest Y^* is also discrete and takes values $\{\Delta_g(\tilde{y}_k) : k \in \{1,\dots K\}$ and $g \in \{0,1\}\}$.

C Regularity conditions for Proposition 3

In this section, we state the regularity conditions that are assumed in Proposition 3. These are Assumptions 1-2 in Moon and Schorfheide (2012) and we restate them here for completeness.

Recall Equation (6) in Section 3.1, which defined the likelihood function of the observed training dataset D_N

$$\mathcal{L}(D_N;\eta) := \prod_{i=1}^{N} \left(\prod_{k=1}^{K} \tilde{\eta}_k(G_i,W_i)^{1\{\tilde{Y}_i=\tilde{y}_k\}} \right) P(G_i,W_i), \qquad l(D_N;\eta) := \log(\mathcal{L}(D_N;\eta)).$$

Define

$$\hat{J}_N := K_N^{-1} \left(-\frac{\partial^2 l(D_N;\eta)}{\partial\tilde{\eta}\partial\tilde{\eta}'} \right) K_N^{-1} \text{ and } s := \hat{J}_N^{-1/2} K_N(\tilde{\eta} - \hat{\tilde{\eta}}_N^{MLE}),$$

where $\hat{\tilde{\eta}}_N^{MLE}$ is the maximum likelihood estimator of $\tilde{\eta}$, and K_N is a deterministic matrix with elements that diverge as $N \to \infty$ and is chosen such that $\hat{J}_N$ is convergent. Let $\pi(s|D_N)$ denote the posterior distribution of the transformed parameter s. Let $\tilde{\eta}_0$ denote the true value of $\tilde{\eta}$ in the population.

Assumption 4 (Assumption 1 of Moon and Schorfheide (2012))**.** *Assume that*

1. *The sequence of maximum likelihood estimators $\hat{\tilde{\eta}}_N^{MLE}$ are consistent. The matrix $\|D_N\| \to \infty$. The likelihood function $cL(D_N;\eta)$ is twice continuously differentiable with probability approaching one such that $\hat{J}_N$ is well-defined. The Hessian of the log-likelihood function l has a positive definite limit: $\hat{J}_N \xrightarrow{d} J_0 > 0$ and $\hat{J}_N^{-1} \xrightarrow{d} J_0^{-1}$.*

2. *The posterior distribution of $\tilde{\eta}$ is asymptotically normal, meaning $\|\pi(s|D_N) - N(0,I)\| \xrightarrow{p} 0$.*

In our application, the assumptions here are simple to check as the model is fully parametric and fits directly into the set-up in Moon and Schorfheide (2012). Let $\pi(\eta^*|\tilde{\eta})$ denote the conditional distribution of η^* given $\tilde{\eta}$ under the prior distribution π. We additionally make the following assumption.

Assumption 5 (Assumption 2 of Moon and Schorfheide (2012)). *Let* $N_\delta(\tilde{\eta}) = \{\tilde{\eta} : \|\tilde{\eta} - \tilde{\eta}_0\| < \delta\}$. *Assume that there exists a* $\delta > 0$ *and constant* $M(\tilde{\eta}_0, \delta)$ *such that* $\|\pi(\eta^*|\tilde{\eta}) - \pi(\eta^*|\tilde{\eta}')\|_{TV} \leq M(\tilde{\eta}_0, \delta)\|\tilde{\eta} - \tilde{\eta}'\|$ *for* $\tilde{\eta}, \tilde{\eta}'$ *in* $N_\delta(\tilde{\eta}_0)$.

D Proofs of Main Results

Proof of Proposition 1

The objective function in the first-best problem is simply an integrated risk function that assigns prior weights $\pi(\eta)$ to the parameter. Standard arguments in statistical decision theory immediately implies that the first-best admissions rule can be obtained by constructing the admissions rule that minimizes posterior expected social welfare at any realization of the training dataset that occurs with positive prior probability. That is, the first-best admissions rule $t^*(g, w; D_N)$ at any training dataset D_N that occurs with positive probability may be obtained by solving

$$\begin{aligned} &\max_{t(g,w)} \sum_{(g,w)} \psi_g \mathbb{E}_{\pi|D_N}\left[\theta^*(g,w)\right] t(g,w) P(g,w) \\ &\text{s.t.} \sum_{(g,w)} t(g,w) P(g,w) \leq C. \end{aligned}$$

The social planner's posterior beliefs are constructed as described in Section 3.1.

Without loss of generality, order groups defined by the characteristics (g, w) using $\psi_g \cdot \mathbb{E}_{\pi|D_N}\left[\theta^*(g,w)\right]$. Let $(g_1, w_1), \ldots, (g_N, w_M)$ denote such an ordering with $M = 2^{J+1}$, where $\psi_j \mathbb{E}_{\pi|D_N}\left[\theta_j^*\right] = \psi_{g_j} \mathbb{E}_{\pi|D_N}\left[\theta^*(g_j, w_j)\right]$ is the j-the element of the ordering and

$$\psi_1 \mathbb{E}_{\pi|D_N}\left[\theta_1^*\right] \geq \psi_2 \mathbb{E}_{\pi|D_N}\left[\theta_2^*\right] \geq \ldots \geq \psi_M \mathbb{E}_{\pi|D_N}\left[\theta_M^*\right].$$

Let $j(C)$ be the largest index of this list such that $\sum_{j \leq j(C)} P_j \leq C$, where $P_j = P(g_j, w_j)$.

If $\sum_{j \leq j(C)} P_j = C$, then the social planner's optimal admissions rule takes the form:

$$t(g_j, w_j) = \begin{cases} 1 & \text{if } j \leq j(C), \\ 0 & \text{otherwise.} \end{cases}$$

Otherwise, the social planner could reallocate admissions probabilities $t(g, w)$ in a manner that strictly raised expected social welfare under her posterior $\pi|D_N$. So, define $\tau(C) = \psi_{j(C)} \mathbb{E}_{\pi|D_N}\left[\theta_{j(C)}^*\right]$ and the social planner's optimal admissions rule can be written as

$$t(g, w) = \mathbb{1}\left\{\mathbb{E}_{\pi|D_N}\left[\theta^*(g,w)\right] > \frac{\tau(C)}{\psi_g}\right\},$$

where the case of ties with $\mathbb{E}_{\pi|D_N}\left[\theta^*(g,w)\right] > \frac{\tau(C)}{\psi_g}$ is handled by setting $t(g, w) = 1$. Defining $\tau^*(g; C) = \frac{\tau(C)}{\psi_g}$ delivers the result for this case.

Next, if $\sum_{j \leq j(C)} P_j < C$, then the social planner's optimal admissions rule takes the form

$$t(g_j, w_j) = \begin{cases} 1 & \text{if } j \leq j(C), \\ C - \sum_{j \leq j(K)} P_j & \text{if } j = j(C) + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Again, otherwise, the social planner could reallocate admissions probabilities $t(g, w)$ in a manner that strictly raised expected social welfare under her posterior $\pi | D_N$. Now, define $\tau^*(C) = \psi_{j(C)+1} \mathbb{E}_{\pi | D_N} \left[\theta^*_{j(C)+1} \right]$. The social planner's optimal admissions rule can again be rewritten as

$$t(g, w) = \mathbb{1} \left\{ \mathbb{E}_{\pi | D_N} \left[\theta^*(g, w) \right] > \frac{\tau(C)}{\psi_g} \right\},$$

where the case of ties with $\mathbb{E}_{\pi | D_N} \left[\theta^*(g, w) \right] = \frac{\tau(C)}{\psi_g}$ is by setting $t(g, w) = C - \sum_{j \leq j(K)} P_j$. The result then follows for this case as well. □

Proof of Proposition 2

We provide one direction of the result and refer the reader to Proposition 1 of Poirier (1998) for the other direction. By Bayes Rule, the marginal posterior for η^* is given by

$$\pi(\eta^* | D_N) = \frac{\pi(\eta^*) \mathcal{L}(D_N | \eta^*)}{\mathcal{L}(D_N)},$$

where $\pi(\eta^*)$ is the marginal prior for η^* and $\mathcal{L}(D_N; \eta^*)$ is the likelihood conditional on η^*, which is obtained noting that by

$$\mathcal{L}(D_N; \tilde{\eta}, \eta^*) = \mathcal{L}(D_N; \tilde{\eta})$$

and computing

$$\mathcal{L}(D_N; \eta^*) = \int_{\tilde{\eta}} \pi(\tilde{\eta} | \eta^*) \mathcal{L}(D_N; \tilde{\eta}) d\tilde{\eta}.$$

Finally, $\mathcal{L}(D_n)$ is the marginal distribution over the training dataset and it is obtained by

$$\mathcal{L}(D_N) = \int_{\eta^*} \int_{\tilde{\eta}} \pi(\tilde{\eta} | \eta^*) \mathcal{L}(D_N; \tilde{\eta}) d\tilde{\eta}.$$

If $\tilde{\eta} \perp\!\!\!\perp \eta$ under π, then $\mathcal{L}(D_N; \eta^*) = \mathcal{L}(D_N)$. The result follows immediately as this implies that $\pi(\eta^* | D_N) = \pi(\eta^*)$. □

Proof of Proposition 3

For simplicity, we additionally denote the total variation distance between two probability measures as $d_{TV}(F, G) = \|F - G\|_{TV}$. Let $\tilde{\eta}^{MLE}$ denote the maximum likelihood estimate of $\tilde{\eta}(g, w)$ and let $\tilde{\eta}_0$ denote the true value of $\tilde{\eta}$ in the population. Applying the triangle inequality, we have that

$$\|\pi(\eta^*|\mathbf{D}_N) - \pi(\eta^*|\hat{f}_N)\|_{TV} = \|\pi(\eta^*|\mathbf{D}_N) - \pi(\eta^*|\tilde{\eta}_N^{MLE}) + \pi(\eta^*|\tilde{\eta}_N^{MLE}) - \pi(\eta^*|\hat{f}_N)\|_{TV}$$

$$\leq \|\pi(\eta^*|\mathbf{D}_N) - \pi(\eta^*|\tilde{\eta}_N^{MLE}(w, g))\|_{TV} + \|\pi(\eta^*|\tilde{\eta}_N^{MLE}) - \pi(\eta^*|\hat{f}_N)\|_{TV},$$

Under the stated regularity conditions in Appendix C, Theorem 1 in Moon and Schorfheide (2012) applies and the first term converges in probability to zero. Therefore, it is sufficient to show that

$$\|\pi(\eta^*|\tilde{\eta}_N^{MLE}) - \pi(\eta^*|\hat{f}_N)\|_{TV} \xrightarrow{p} 0.$$

To do so, define the sequence of events $A_n = \{\|\tilde{\eta}_N^{MLE} - \tilde{\eta}_0\| < \delta, \|\hat{f}_N - \tilde{\eta}_0\| < \delta\}$. The probability of these events goes to one as $N \to \infty$ as both the MLE estimator and the algorithm's prediction function are consistent. We place ourselves on these events without loss of generality. On these events, we apply the Lipschitz condition to show that

$$\|\pi(\eta^*|\tilde{\eta}_N^{MLE}) - \pi(\eta^*|\hat{f}_N)\|_{TV} \leq M(\tilde{\eta}_0, \delta)\|\tilde{\eta}_N^{MLE} - \hat{f}_N\|,$$

where $\|\tilde{\eta}_N^{MLE} - \hat{f}_N\| \xrightarrow{p} 0$ because both are consistent. Therefore, we conclude

$$\|\pi(\eta^*|\tilde{\eta}_N^{MLE}) - \pi(\eta^*|\hat{f}_N)\|_{TV} \xrightarrow{p} 0,$$

establishing the result. □

Proof of Proposition 4

Let $\mathcal{M} = \{m : \tilde{m} \subseteq m\}$. We prove this result in steps:

Step 1: We show that for any model $m \in \mathcal{M}$, the non-discriminatory decision-maker constructs the same rank-ordering over the population and therefore, for fixed capacity constraint C, she selects the same admissions rule across these models.

To see this, consider any such m. If the human decision-maker is allowed to select decision rules that use group membership G, then she chooses her admissions rule to maximize

$$\sum_{g\in\{0,1\}} \sum_{w_m\in\{0,1\}^{|m|}} \left\{ \sum_{w_{-m}\in\{0,1\}^{J-|m|}} \theta^*_{\pi_{\tilde{m}}}(w_m, w_{-m}) P(g, w_m, w_{-m}) \right\} t(g, w_m)$$

$$= \sum_{g\in\{0,1\}} \sum_{w_{\tilde{m}}\in\{0,1\}^{|\tilde{m}|}} \left\{ \sum_{w_{m-\tilde{m}}\in\{0,1\}^{|m|-|\tilde{m}|}} \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) t(g, w_{\tilde{m}}, w_{m-\tilde{m}}) P(g, w_{\tilde{m}}, w_{m-\tilde{m}}) \right\},$$

where $P(g, w_{\tilde{m}}, w_{m-\tilde{m}}) = \sum_{w_{-\tilde{m}} \in \{0,1\}^{J-|m|}} P(g, w_{\tilde{m}}, w_{m-\tilde{m}}, w_{-\tilde{m}})$. Therefore, the human decision-maker divides the population into groups divided by the characteristics W_m, G and rank orders the groups using $\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}})$. Any groups with the same value of $w_{\tilde{m}}$ are given the same ranking, which is the same ranking as if the social planner only allowed the human decision-maker to use model $\tilde{m}$. Because the rankings are the same, for fixed capacity C, the admissions rules are equivalent between model $\tilde{m}$ and model m based upon Proposition 1.

Similarly, if the social planner bans the human decision-maker from using group membership G, then the human decision-maker chooses her admissions rule to maximize

$$\sum_{w_{\tilde{m}} \in \{0,1\}^{|\tilde{m}|}} \left\{ \sum_{w_{m-\tilde{m}} \in \{0,1\}^{|m|-|\tilde{m}|}} \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) t(w_{\tilde{m}}, w_{m-\tilde{m}}) P(w_{\tilde{m}}, w_{m-\tilde{m}}) \right\},$$

where $P(w_{\tilde{m}}, w_{m-\tilde{m}}) = P(0, w_{\tilde{m}}, w_{m-\tilde{m}}) + P(1, w_{\tilde{m}}, w_{m-\tilde{m}})$. Again, the human decision-maker divides the population into groups based upon the characteristics W_m and rank orders the groups using $\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}})$. Because $\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}})$ does not vary across group membership G and neither does the non-discriminators preferences, this is the same ranking as if the human decision-maker could use G in her admissions rule. Once again, it implies that the admissions rules are equivalent.

Therefore, we conclude that for fixed capacity constraint C, the admissions rules for all models m satisfying $\tilde{m} \subseteq m$ are equivalent. This implies that the social planner is indifferent between these models. For the remainder of the proof, we therefore focus attention on the model $\tilde{m}$ without loss of generality.

Step 2: Consider a model $m \subset \tilde{m}$. If the social planner strictly prefers model m to model $\tilde{m}$, then there exists some pairs $(g, w_{\tilde{m}})$, $(g', w'_{\tilde{m}})$ such that the non-discriminator ranks these pairs differently than the social planner at model $\tilde{m}$ but ranks them in accordance with the social planner's ranking at model m. This cannot occur because the social planner's preferences are aligned with the non-discriminator's preferences, and so they select the same rank-ordering at all models. By a similar argument, we can show that the same is true of any model $m \not\subset \tilde{m}$ with $m \cap \tilde{m} \subset \tilde{m}$ as well. □

Proof of Proposition 5

Consider a discriminatory human decision-maker at model $\tilde{m}$. If she cannot use group membership, she selects an admissions rule to maximize

$$\sum_{w_{\tilde{m}}} \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) \left\{ P(0|w_{\tilde{m}}) + \bar{\lambda}_1 P(1|w_{\tilde{m}}) \right\} t(w_{\tilde{m}}) P(w_{\tilde{m}}).$$

Defining $\bar{\lambda}(w_{\tilde{m}}) = P(0|w_{\tilde{m}}) + \bar{\lambda}_1 P(1|w_{\tilde{m}})$, the discriminatory human decision-maker ranks the population according to $\bar{\lambda}(w_{\tilde{m}}) \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}})$.

We show that at model controls $\tilde{m}$ with group membership banned, the discriminatory human decision-maker ranks the population in the same manner as the non-

discriminatory human decision-maker. That is,

$$\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) = \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}) \implies \bar{\lambda}(w_{\tilde{m}})\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) = \bar{\lambda}(w'_{\tilde{m}})\theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}})$$
$$\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}) \implies \bar{\lambda}(w_{\tilde{m}})\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \bar{\lambda}(w'_{\tilde{m}})\theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}).$$

First, consider the case with $\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) = \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}})$. By the relevance assumption, $w_{\tilde{m}} = w'_{\tilde{m}}$ and therefore, $\bar{\lambda}(w_{\tilde{m}}) = \bar{\lambda}(w'_{\tilde{m}})$. The result follows.

Second, consider the case $\theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}})$. The disadvantage condition gives that $\frac{P(0,w_{\tilde{m}})}{P(1,w_{\tilde{m}})} > \frac{P(0,w'_{\tilde{m}})}{P(1,w'_{\tilde{m}})}$, and so by Bayes' rule $\frac{P(0|w_{\tilde{m}})}{P(1|w_{\tilde{m}})} > \frac{P(0|w'_{\tilde{m}})}{P(1|w'_{\tilde{m}})}$. Since $P(0|w_{\tilde{m}}) = 1 - P(1|w_{\tilde{m}})$, this inequality implies that

$$P(1|w_{\tilde{m}}) < P(1|w'_{\tilde{m}}) \text{ and } \bar{\lambda}(w_{\tilde{m}}) > \bar{\lambda}(w'_{\tilde{m}}).$$

Therefore, at model controls $\tilde{m}$ with group membership banned, the social planner implements her preferred rank ordering and achieves the first-best outcome. □

Proof of Proposition 6

At any cutoff C with group membership banned, the discriminatory human decision-maker wishes to select an admissions rule to maximize

$$\begin{aligned} U(t;\bar{\lambda}) &= \sum_{w\in\{0,1\}^J} \left(P(0|w) + \bar{\lambda}_1 P(1|w)\right) \tilde{\theta}_{\pi_{\tilde{m}}}(w) t(w) P(w) \\ &= \sum_{w\in\{0,1\}^J} \bar{\lambda}(w)\tilde{\theta}_{\pi_{\tilde{m}}}(w) t(w) P(w), \text{ where } \bar{\lambda}(w) = P(0|w) + \bar{\lambda}_1 P(1|w). \end{aligned}$$

Therefore, if $P(G = 1|W = w) \neq P(G = 1|W = w')$ for all $w, w' \in \{0,1\}^J$ with $w \neq w'$, then $\bar{\lambda}(w)$ varies across the population. That is, for $w = (w_{\tilde{m}}, w_{-\tilde{m}})$, $w' = (w_{\tilde{m}}, w'_{-\tilde{m}})$, it may be the case that

$$\bar{\lambda}(w)\tilde{\theta}_{\pi_{\tilde{m}}}(w) \neq \bar{\lambda}(w')\tilde{\theta}_{\pi_{\tilde{m}}}(w'),$$

even though $\tilde{\theta}_{\pi_{\tilde{m}}}(w) = \tilde{\theta}_{\pi_{\tilde{m}}}(w')$. It is immediate that the discriminatory firm wishes to rank order the population using $\bar{\lambda}_g(w)\tilde{\theta}_{\pi_{\tilde{m}}}(w)$, even though her model for the outcome of interest is simply $\tilde{m}$. □

Proof of Proposition 7

Consider the social planner's objective function evaluated at model regulations m and re-write it as

$$\int_C \left(\sum_{w \in \{0,1\}^J} \mathbb{E}_\eta \left[\tilde{\theta}_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{\lambda,C}(w;m) \right] P(w) \right) h(C) dC$$

$$= \int_C \left(\sum_{w \in \{0,1\}^J} \left\{ \sum_{\tilde{m}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{ND,C}(w;m) \eta(\tilde{m}, ND) + \sum_{\tilde{m}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{D,C}(w;m) \eta(\tilde{m}, D) \right\} P(w) \right) h(C) dC,$$

where $t^{\tilde{m}}_{ND,C}(g,w;m)$ is the admissions rule selected by a non-discriminatory human decision-maker at true model $\tilde{m}$ and cutoff C and $t^{\tilde{m}}_{D,C}(g,w;m)$ is defined analogously for the discriminatory human decision-maker. We next add and subtract the social planner's payoff at her first-best admissions rule

$$\int_C \left(\sum_{w \in \{0,1\}^J} \left\{ \sum_{\tilde{m}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{*,C}(w) \eta(\tilde{m}, ND) + \sum_{\tilde{m}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) t^{\tilde{m}}_{*,C}(w) \eta(\tilde{m}, D) \right\} P(w) \right) h(C) dC,$$

where $t^{\tilde{m}}_{*,C}(g,w)$ is the social planner's optimal admissions rule at true model $\tilde{m}$ and cutoff C. This is a constant, and so it does not affect the optimizer. Maximizing the original objective is equivalent to maximizing

$$\int_C \left(\sum_{w \in \{0,1\}^J} \left\{ \sum_{\tilde{m}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{ND,C}(w;m) - t^{\tilde{m}}_{*,C}(w) \right] \eta(\tilde{m}, ND) P(w) \right\} \right) h(C) dC)$$

$$+ \int_C \left(\sum_{w \in \{0,1\}^J} \left\{ \sum_{\tilde{m}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{D,C}(w;m) - t^{\tilde{m}}_{*,C}(w) \right] \eta(\tilde{m}, D) P(w) \right\} \right) h(C) dC)$$

$$= \int_C \left(\sum_{\tilde{m}} \left\{ \sum_{w \in \{0,1\}^J} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{ND,C}(w;m) - t^{\tilde{m}}_{*,C}(w) \right] P(w) \right\} \eta(ND|\tilde{m}) \eta(\tilde{m}) \right) h(C) dC$$

$$+ \int_C \left(\sum_{\tilde{m}} \left\{ \sum_{w \in \{0,1\}^J} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{D,C}(w;m) - t^{\tilde{m}}_{*,C}(w) \right] P(w) \right\} \eta(D|\tilde{m}) \eta(\tilde{m}) \right) h(C) dC.$$

Using the fact that $\eta(ND|\tilde{m}) = 1 - \eta(D|\tilde{m})$, this becomes

$$\int_C \left(\sum_{\tilde{m}} \left\{ \sum_{w \in \{0,1\}^J} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{ND,C}(w;m) - t^{\tilde{m}}_{*,C}(w) \right] P(w) \right\} \eta(\tilde{m}) \right) h(C) dC$$
$$+ \eta(D) \int_C \left(\sum_{\tilde{m}} \left\{ \sum_{w \in \{0,1\}^{J+1}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{D,C}(w;m) - t^{\tilde{m}}_{ND,C}(w;m) \right] P(w) \right\} \eta(\tilde{m}|D) \right) h(C) dC$$

Flipping the sign, maximizing the social welfare function is equivalent to minimizing

$$\int_C \left(\sum_{\tilde{m}} \left\{ \sum_{w \in \{0,1\}^J} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{*,C}(w) - t^{\tilde{m}}_{ND,C}(w;m) \right] P(w) \right\} \eta(\tilde{m}) \right) h(C) dC$$
$$+ \int_C \left(\sum_{\tilde{m}} \left\{ \sum_{w \in \{0,1\}^{J+1}} \tilde{\theta}_{\pi_{\tilde{m}}}(w) \left[t^{\tilde{m}}_{ND,C}(w;m) - t^{\tilde{m}}_{D,C}(w;m) \right] P(w) \right\} \eta(\tilde{m}|D) \right) h(C) dC \eta(D)$$
$$= \sum_{\tilde{m}} \left\{ \int_C \mathbb{E} \left[\tilde{\theta}_{\pi_{\tilde{m}}}(w) \left(t^{\tilde{m}}_{*,C}(w) - t^{\tilde{m}}_{ND,C}(w;m) \right) \right] h(C) dC \right\} \eta(\tilde{m})$$
$$+ \eta(D) \sum_{\tilde{m}} \left\{ \int_C \mathbb{E} \left[\tilde{\theta}_{\pi_{\tilde{m}}}(w) \left(t^{\tilde{m}}_{ND,C}(w;m) - t^{\tilde{m}}_{D,C}(w;m) \right) \right] h(C) dC \right\} \eta(\tilde{m}|D).$$

□

Proof of Proposition 8

Suppose, for sake of contradiction, that the equilibrium level of discrimination was zero. This means that for all beliefs $\pi_{\tilde{m}}$ and capacity constraints C

$$t^{\tilde{m}}_{\bar{\lambda},C}(m^*) = t^{\tilde{m}}_{(1,1),C}(m^*).$$

First, suppose that group membership is not banned at m^*. By the stated assumption, there exists a pair of characteristics w_{m^*}, w'_{m^*} such that

$$\theta^*_{\pi_{m^*}}(w_{m^*}) > \theta^*_{\pi_{m^*}}(w'_{m^*})$$
$$\theta^*_{\pi_{m^*}}(w'_{m^*}) > \bar{\lambda}_1 \theta^*_{\pi_{m^*}}(w_{m^*}).$$

Since the distribution over capacity constraints has full support, this implies that there exists values of C that occur with positive probability such that $t^{m^*}_{\bar{\lambda},C}(m^*) \neq t^{m^*}_{(1,1),C}(m^*)$ as w'_{m^*} is admitted before w_{m^*} by the discriminators but not by the non-discriminators.

Next, suppose that group membership is banned at m^*. Then, the non-discriminatory human decision-makers with beliefs π_{m^*} rank-order according to $\theta^*_{\tilde{\pi}}(w_{m^*})$ and discriminatory human decision-makers with beliefs π_{m^*} rank-order according to $\bar{\lambda}(w_{m^*})\theta^*_{\tilde{\pi}}(w_{m^*})$.

Again, the stated assumption, there exists a pair of characteristics w_{m^*}, w'_{m^*} such that

$$\theta^*_{\pi_{m^*}}(w_{m^*}) > \theta^*_{\pi_{m^*}}(w'_{m^*})$$
$$\bar{\lambda}(w'_{m^*})\theta^*_{\pi_{m^*}}(w'_{m^*}) > \bar{\lambda}(w_{m^*})\theta^*_{\pi_{m^*}}(w_{m^*}).$$

The contradiction proceeds as before. □

Proof of Proposition 10

Recall in Step 1 in the proof of Proposition 5, we show that the rank-ordering used by the discriminatory human decision-maker is the same as a non-discriminatory human decision-maker with ground truth model $\tilde{m}$ if the social planner implements model controls $m = \tilde{m}$ and bans group membership.

Therefore, at the model regulations $m^*(\tilde{m}) = \tilde{m}$ with group membership banned, all discriminatory human decision-makers and non-discriminatory human decision-makers with the same ground-truth model select the same rank ordering. This immediately implies $t^{\tilde{m}}_{\bar{\lambda},C}(m^*) = t^{\tilde{m}}_{(1,1),C}(m^*)$ as all human decision-makers simply admit individuals according to their chosen rank-ordering until the capacity constraint is satisfied. □