An Economic Approach to Regulating Algorithms*

Ashesh Rambachan†  Jon Kleinberg‡  Jens Ludwig§  Sendhil Mullainathan¶

February 16, 2020

Abstract

There is growing concern of "algorithmic bias" - that algorithms used in decision-making might bake in or exaggerate discrimination in society. When will these biases arise? What should be done about them? Such questions at first blush appear to be best addressed using computational frameworks. We argue that they are naturally answered using the tools of welfare economics - the standard framework applied to address many other policy problems including discrimination by people. Our framework incorporates the novel elements that algorithms introduce into the decision-making process and fully models the welfare function of a policy-maker as well as the potential incentives of the algorithm designer. Core to our model is that the standard supervised learning framework incorporates two distinct algorithms - the training algorithm that produces a prediction function and a decision rule based on these predictions. Our first core result is an irrelevance result - the equity preferences of the social planner have no effect on the training procedure. For example, they would allow use of all group membership variables. We then turn to the situation where a (possibly discriminatory) private actor chooses the algorithm rather than the social planner. Absent algorithms, optimal regulation in our framework resembles existing policy, consisting of prohibitions against disparate treatment and tests for disparate impact. Optimal algorithmic regulation, on the other hand, depends crucially on the disclosure regime. If decision-makers must reveal all underlying inputs (the data, training procedure and decision rule), once again a strong result arises: it is optimal to let any characteristic that is predictive of the outcome of interest be used. Absent such disclosure, little is gained relative to human decision-making. Importantly, under full disclosure, the presence of algorithms strictly reduces discrimination relative to a world in which humans make all the decisions. Our results show how a welfare-economics approach can produce policy implications quite distinct from the common wisdom in more computational approaches and direct attention to modifications of our assumptions under which these irrelevance results break.

*We thank 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 for valuable feedback and Joshua Gans for useful comments as a conference discussant. Rambachan gratefully acknowledges financial support from the NSF Graduate Research Fellowship (Grant DGE1745303). All errors are our own.
†Harvard University, Department of Economics: asheshr@g.harvard.edu
‡Cornell University, Department of Computer Science and Department of Information Science: kleinber@cs.cornell.edu
§University of Chicago, Harris School of Public Policy and NBER: jlldudwig@uchicago.edu
¶University of Chicago, Booth School of Business and NBER: Sendhil.Mullainathan@chicagobooth.edu
1 Introduction

The growing use of algorithms to inform key decisions such as hiring, credit approvals, pre-trial release and medical testing has been accompanied by growing concerns about the risk of “algorithmic bias.” Though there is no single definition, algorithmic bias is often used as a blanket term surrounding fears that algorithmic decision-making may reflect, or even exacerbate existing discrimination in society.\(^1\) Such concerns arise primarily because the underlying historical data that these algorithms are trained upon might reflect past prejudice and discrimination found in society itself. For example, resume screening software is trained upon past, potentially discriminatory hiring decisions. Criminal records used in recidivism prediction may bake in police and judicial biases. A large body of research empirically documents such biases. A theoretical literature, largely in computer science, explores how such biases arise and how their presence should affect the design of algorithms.

Concerns about algorithmic bias are at their heart questions of optimal policy. However, existing research misses three key ingredients that economists typically take to be central to any policy analysis. First, there is rarely a full specification of a social planner’s preferences over outcomes from which policy can be derived; instead some definition of fairness is viewed as an axiomatic constraint on the algorithm. Second, there is rarely a complete description of the full suite of policy tools available to influence outcomes; instead the focus is solely on the design of the algorithm. Finally, incentive problems are often overlooked; instead it is assumed that the builder of the algorithm shares society’s preferences.

In this paper, we analyze policy questions surrounding algorithmic bias by incorporating the novelties of supervised machine learning algorithms into a canonical welfare economics framework. We focus on cases where empirically-based supervised machine learning algorithms inform decision-making in screening decisions. In a screening decision, a decision-maker must select one or more people from a larger pool based upon a prediction of an uncertain outcome of interest. The decision-maker may construct a supervised machine learning algorithm using training data, which consists of measured characteristics and outcomes for some sample of individuals. We model a supervised machine learning algorithm as consisting of two components: a “predictive algorithm,” which takes in training data and returns a prediction function, and a “decision rule,” which uses the constructed prediction

\(^1\)The literature on algorithmic bias is vast. A curated list of papers on algorithmic bias is assembled at https://www.fatml.org/resources/relevant-scholarship. Barocas et al. (2019) provides a textbook introduction to the computer science literature on this topic and Cowgill and Tucker (2019) provides a survey for economists. We discuss the connections between this paper and this large literature at the end of the introduction and in Section 3.
function to make decisions. A policymaker may therefore regulate supervised machine learning algorithms by regulating the predictive algorithm or the decision rule. Next, we embed concerns about algorithmic bias within a social welfare function, which summarizes society’s preferences over the resulting outcomes of the screening problem. The social welfare function may contain an explicit preference for more equitable outcomes across groups. We take the social welfare function as a primitive and derive its implications for the construction of supervised machine learning algorithms across two policy environments.

With this model of algorithmic decision-making in hand, we establish a series of sharp null results, many of which run counter to the prevailing wisdom around how best to manage algorithmic bias. Even for those who do not believe these counter-intuitive results, the model provides a strong baseline against which must be asked: what assumptions of the model must be modified to break these results?

We first consider the algorithm choice of a social planner that wishes to maximize social welfare and makes the screening decisions herself. This is the social planner’s “first-best problem.” We show that even a social planner with explicit equity preferences wishes to make use of data that may bake-in historical biases. The intuition underlying this result is exceedingly simple. In the supervised machine learning pipeline, the predictive algorithm simply summarizes information in the observed training data and given that information, the social planner may modify the decision rule to maximize social welfare. Irrespective of her preferences, the social planner does not wish to destroy potentially useful information. So long as the social planner believes the data contain useful signal about the outcome of interest, she will make use of that signal.

This is an equity irrelevance result: explicit equity motives have no effect on the predictive algorithm. In constructing predictions, the social planner’s only desired property is accuracy. Equity preferences only modify the decision rule, which takes the constructed prediction function as an input. This implies that all characteristics, including group membership, are given to the predictive algorithm, and the social planner does not wish to place any additional constraints on her training procedure. Moreover, our equity irrelevance result is robust to a wide variety of common concerns surrounding sources of algorithmic bias. For example, it holds even if the observed outcome in the training data differs from the outcome of interest, there are differences in the conditional base rates of the outcome of interest across groups or there are differences in the distribution of characteristics across groups. The “first-best problem” is relevant for applications where policymakers construct the predictive algorithm and make the screening decisions themselves.
In many other cases, however, the social planner neither makes the screening decisions nor constructs the algorithm - private actors do and some of them may even wish to discriminate against the disadvantaged group. For example, individual firms make hiring decisions and choices around how to use algorithms in the hiring process. In these cases, the social planner faces a regulation problem, in which she must use the tools available to define discriminatory behavior and to detect it when it is happening. We model the social planner as regulating which characteristics may be used in the decision rules of private actors. We assume some private actors are taste-based discriminators, there are no average group differences in the outcome of interest conditional on observable characteristics and there are no additional affirmative action preferences. To build intuition, we consider optimal regulation absent algorithms. We show that optimal regulation in our framework resembles existing policy, consisting of prohibitions against disparate treatment and tests for disparate impact. Intuitively, the social planner wishes to ensure that discriminators select similar decision rules as non-discriminators and optimal regulation involves a “flexibility tradeoff” – allowing more characteristics leads to more accurate predictions, but extra characteristics may also be used to screen out members of the disadvantaged group.

We then consider the case where these private actors use an algorithm in their screening decisions. Optimal regulation changes substantially as long as disclosure of all parts of the algorithm – both the training algorithm and the decision rule. We refer to such disclosures as an algorithmic audit (Kleinberg et al., 2018). With such algorithmic audits in place, there is no longer a prohibition against disparate treatment or a test for disparate impact. Instead, it is optimal to allow the algorithm to have access to all characteristics. Furthermore, we crucially show that the equilibrium level of discrimination declines if the social planner may carry out such algorithmic audits. The key insight is that algorithmic audits enable the social planner to enforce that all human decision-makers (whether they are discriminatory or non-discriminatory) select the same admissions rule if they have the same prediction function. With the correct regulatory system in place – specifically, one that allows algorithmic audits to be conducted – the introduction of predictive algorithms into screening decisions can lead to not only improved prediction, but can simultaneously make it easier to detect discrimination in the market.

To highlight the importance of algorithmic audits, we finally consider optimal regulation if the social planner can only compel human decision-makers to disclose their decision rules, but not the predictive algorithm. In this case, even though the human decision-makers rely on algorithmic pre-
dictions, optimal regulation is the same as with a human-only decision loop. The social planner treats the predictive algorithm as a “black box” and focuses only on the revealed decision rules, applying the same disparate impact criteria. Therefore, to realize the gains of algorithmic decision-making, the entire algorithmic pipeline must be disclosed to the social planner.

Our approach is crucially different from that taken by a large community of researchers in computer science and statistics. Existing research typically begins by noting that a supervised machine learning algorithm simply generates a mapping from observed data into predictions or decisions and then formally defines what it means for such a mapping 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. 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 start by defining fairness in terms of preferences over the resulting outcomes of the screening decision using a social welfare function. Rather than deriving fairness as a given property of prediction functions or decision rules, we take the preferences that are summarized by the social welfare function as our primitive and derive its implications for algorithm construction. Moreover, existing research in computer science and statistics tends to only consider the first-best problem in which a benevolent social planner controls the design and implementation of the algorithm. It overlooks agency problems that may arise when algorithms are designed and implemented by third-party decision-makers. Our analysis of the second-best problem is a new contribution to the computer science literature, highlighting the value of an economic perspective to this problem. We discuss the connections to the literature in computer science in more detail in Section 3.

Finally, we highlight two recent papers in economics that also incorporate insights from microeconomic theory 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 human decision-makers are aware that predictions are manipulated by a planner, then human decision-makers may optimally ignore these predictions in their decisions.

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 labels \( Y^* \in \{0, 1\}, \tilde{Y} \in \{0, 1\} \), where the label \( Y^* \) is the “outcome of interest” and the label \( \tilde{Y} \) is the “measured outcome.” The population of individuals is summarized by a joint distribution \( P \) over the random vector \((Y^*, \tilde{Y}, G, W)\).

Let
\[
P(g, w) = \mathbb{P} \{G = g, W = w\}, \quad P(w) = \mathbb{P} \{W = w\}
\]
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
\[
\theta^*(g, w) = \mathbb{E} [Y^* | G = g, W = w], \quad \tilde{\theta}(g, w) = \mathbb{E} [\tilde{Y} | G = g, W = w]
\]
denote the average outcome of interest \( Y^* \) and the average measured outcome \( \tilde{Y} \) among individuals that belong to group \( g \) with characteristics \( w \).

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. Admissions decisions are made based on 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

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

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,
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 $(Y^*, \bar{Y}, G, W)$ 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 characteristics 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:

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

where $\psi_g \geq 0$ are generalized social welfare weights placed upon individuals in group $g$ (Saez and Stantcheva, 2016). 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. This 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.”

**Example 1** (continuing from p. 7). Productive workers \( Y^* = 1 \) produce output if hired and the social welfare function depends on total output. However, the social planner wishes to protect African-American 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. 7). Loans are given out to consumers and credit-worthy borrowers \( Y^* = 1 \) will not 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.

**Remark 1.** In Section A of the Appendix, we provide a simple motivation for the social welfare function in Equation 4. 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 if the utility of an individual from the disadvantaged group is uniformly lower than the utility of an individual from the advantaged group. We interpret this as capturing unmodeled discrimination against the disadvantaged group or existing disparities across groups in a reduced form manner.

### 2.4 The training dataset

From the social welfare function in Equation (4), 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, we assume that the average outcome of interest \( \theta^*(g, w) \) is not known, and therefore 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. Notice that 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.

Finally, we return to our two earlier examples to briefly discuss examples of such training datasets.

**Example 1** *(continuing from p. 7)*. 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. 7)*. 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 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. We show that for a fixed training dataset \(D_N\), the optimal algorithm consists of an optimal decision rule, which is a threshold rule with group-specific thresholds for admission, and an optimal predictive algorithm, which summarizes the social planner’s posterior beliefs about the average outcome of interest \(\theta^*(g, w)\). In other words, the social planner 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, we show that 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 \(\bar{\theta}(g, w)\) and then applying an ex-post adjustment based on her prior beliefs. 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 \(\hat{Y}\) 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 \(y_m^*, \bar{y}_k \in \{0, 1\}\), define the parameters

\[
\mathbb{P} \{Y^* = y_m^*, \hat{Y} = \bar{y}_k | G = g, W = w\} = \eta_{m,k}(g, w) \quad (5)
\]

\[
\mathbb{P} \{Y^* = y_m^* | G = g, W = w\} = \eta_{m,0}(g, w) + \eta_{m,1}(g, w) \equiv \eta_m^*(g, w) \quad (6)
\]

\[
\mathbb{P} \{\hat{Y} = \bar{y}_k | G = g, W = w\} = \eta_{0,k}(g, w) + \eta_{1,k}(g, w) \equiv \tilde{\eta}_k(g, w). \quad (7)
\]

Connecting to our previous notation, notice that \(\eta_{1}^*(g, w) = \theta^*(g, w)\) and \(\tilde{\eta}_1(g, w) = \tilde{\theta}(g, w)\). Let \(\eta = \{\eta_{m,k}(g, w) : m \in \{0, 1\}, k \in \{0, 1\}, g \in \{0, 1\}, w \in \{0, 1\}\}^J\) 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 parameters \(\eta\).

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

\[
\mathcal{L}(D_N; \eta) = \prod_{i=1}^N \tilde{\eta}_1(G_i, W_i)^{\hat{Y}_i} \times \eta_0(G_i, W_i)^{1-\hat{Y}_i} \times P(G_i, W_i). \quad (8)
\]

Since the marginal distribution of \((G, W)\) is known, the likelihood function only depends on the observed training dataset \(D_N\) and the parameters \(\eta\) 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 \(\pi\), 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 \(\pi\), social welfare weights \((\psi_0, \psi_1)\) and capacity constraint \(C\), the social plan-
The social planner’s first-best problem is

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

s.t. $$\sum_{(g, w)} t(g, w; D_N) P(g, w) \leq C$$ 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 $$\pi$$ to average over possible realizations of the training dataset and the parameter $$\eta$$. The capacity constraint must hold at all realizations of the training dataset that occur with positive probability.

We now show that 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 of 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

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

where ties with $$\mathbb{E}_{\pi|D_N} [\theta^*(g, w)] = \tau^*(g; C)$$ 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, and it is the social planner’s optimal predictive algorithm. Moreover, 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.

**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.
Remark 2. If the social welfare weights $\psi$ vary not only across groups but also across other observable features in $W$, Proposition 1 generalizes naturally. In this case, the first-best algorithm still uses a threshold rule, but the admissions thresholds vary based upon the 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 mis-measured 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 $\pi$, $\tilde{\eta}$ is statistically independent of $\eta$.

This result follows directly from Proposition 1 of Poirier (1998). Intuitively, the likelihood function in Equation (8) only depends on the parameter $\eta$ through $\tilde{\eta}(g, w)$. 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 $\tilde{\eta}(g, w)$.

The result in Proposition 2 is quite strong. It 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. This is an incredibly weak condition. It allows the social planner to believe that the measured outcome $\tilde{Y}$ is mis-measured, 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.

---

2In other words, the likelihood function in Equation (8) 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.
3.3 Algorithmic decision-making and the first-best admissions rule

Proposition 1 showed that for a fixed sample size of the training dataset $N$, the social planner’s optimal predictive algorithm constructs a rank-ordering of the population based upon her posterior beliefs $\pi|D_N$. 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 $\pi$ to map these into predictions of the outcome of interest $Y^*$. We now provide a result to show that this intuition is valid in a particular sense asymptotically as the size of the training dataset $N$ grows large.

To develop this result, we first provide a simple definition of an algorithm.

**Definition 3.** An algorithm $A$ is a function that maps a training dataset $D_N$ to a prediction function $A(D_N) = \hat{f}_N$, where $\hat{f}_N : \{0, 1\}^{J+1} \rightarrow [0, 1]$.

An algorithm 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 $\tilde{Y}$. The choice of algorithm may refer to the choice between a variety of different supervised learning algorithms.

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

$$\hat{f}_N(g, w) \overset{p}{\rightarrow} \mathbb{E} [\tilde{Y} | G = g, W = w] \quad \forall (g, w) \in \{0, 1\}^{J+1}.$$  

We now show that 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 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 $\eta^*$ given the parameters $\tilde{\eta}$. The social planner’s posterior beliefs about $\eta^*$ are asymptotically equivalent to the beliefs she would have if she plugged in the predictions of a consistent algorithm into her conditional prior beliefs $\pi(\eta^* | \tilde{\eta})$.

**Proposition 3.** Let $A$ be a consistent algorithm, and assume that the regularity conditions in Section B 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

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

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 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. For this result to hold, we introduce two high-level regularity conditions that are satisfied in the model, which are the same regularity conditions of Moon and Schorfheide (2012). We restate them in Section B of the Appendix 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. In particular, these results imply that the factor in the social planner’s choice of predictive algorithm is accuracy. This implies that 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 constraints that may worsen predictive accuracy. The social planner simply wishes to construct an accurate prediction function of the measured outcome $\tilde{Y}$ using the characteristics $W$ and group membership $G$. Given this estimated prediction function, she maps it into predictions of the outcome of interest using her prior beliefs $\pi$, adjusting the admissions threshold 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. Fairness is defined as an additional constraint that is added in the training procedure of an algorithm, and so 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. 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, and placed no restrictions on the admissions rule, searching 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., see Raghavan et al. (2017), Chouldechova (2017) and Pleiss et al. (2017)). Second, in practice, prediction functions that satisfy a particular definition of predictive fairness may produce downstream, unequal outcomes. Given that our preferences for fairness are

---

3This 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.

4For 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.
ultimately defined over outcomes, it is conceptually attractive to directly summarize these preferences as a social welfare function.

Finally, our result in Proposition 1 is most closely related to several recent papers in computer science and economics. Corbett-Davies et al. (2017) show the optimal prediction functions that satisfy certain definitions of fairness take the form of a threshold classifier with group-specific thresholds for classification.\(^5\) Similarly, Lipton et al. (2018) and Menon and Williamson (2018) provide similar results by characterizing the solution 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 (4). Also taking a social welfare approach, 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. 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 (4) 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. To our knowledge, Proposition 2 and Proposition 3 are novel relative to existing work on algorithmic fairness in computer science and statistics.

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. For example, consider resume screening software. The social planner neither constructs the resume screening software nor can it exactly dictate the hiring decisions of firms.\(^6\) These types of examples 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 influ-

---

\(^5\)These fairness definitions are “statistical parity”, “conditional statistical parity” and “predictive equality.” See Corbett-Davies et al. (2017) for details.

\(^6\)Similarly, credit scores are constructed by private firms using large datasets and used by commercial lenders to make credit approval decisions. The social planner can neither directly control the construct of credit scores nor dictate particular credit approval decisions.
ence their choices. Throughout this section, we refer to the third-party decision-maker as a human decision-maker.

In this section, we 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. That is, in this section we assume an entirely human-driven decision-making pipeline. In the next section, we discuss whether and how the introduction of the algorithm changes the regulatory problem. 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. Because preferences are not aligned, the social planner uses policy instruments to influence the admissions decisions.

We assume that the social planner faces two constraints. First, she faces a policy constraint as she may only influence admissions decisions by placing restrictions on the characteristics that may be used in decision rules. We refer to these restrictions as “model regulations.” This 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. Second, the social planner faces an information constraint as she does not know which human decision-makers are discriminatory and knows less about which characteristics are useful in predicting the outcome of interest than the human decision-makers.

We first show that this model of regulating human decision-makers captures many of our existing intuitions about regulating discrimination. We show that if the social planner only faced discriminatory human decision-makers, then she would ban the use of group membership in admissions decisions. This is analogous to a disparate treatment test. When faced with both non-discriminatory and discriminatory human decision-makers, the social planer faces a flexibility tradeoff – allowing the human decision-makers to use more characteristics leads to more accurate predictions but discriminatory human decision-makers will also use these extra characteristics to screen out members of the disadvantaged group. Therefore, to decide whether an additional characteristic may be used, the social planner must ask whether it is sufficiently predictive of the outcome of interest relative to

---

7 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).

8 This type of flexibility tradeoff commonly arises in “delegation problems,” in which a principal delegates an action to an agent that may have different preferences (“mis-aligned”) than the principal (Holmstrom, 1977, 1984; Alonso and Matouschek, 2008; Amador and Bagwell, 2013).
group membership, mirroring a disparate impact test. Finally, we show that the equilibrium level of discrimination in this purely human-driven decision-making environment (without algorithms) is strictly positive, highlighting the difficulty of detecting discrimination in practice.

4.1 The market of human decision-makers

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

The preferences $\lambda = (\lambda_0, \lambda_1)$ of the human decision-maker 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

$$U(t; \lambda) = \sum_{(g, w) \in \{0, 1\}^{J+1}} \lambda_0^t \theta^t(g, w) t(g, w) P(g, w),$$

where $\lambda = (\lambda_0, \lambda_1)$ 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. For simplicity, we assume there are only two types of preferences $\lambda$ in the market: non-discriminators with $\lambda = (1, 1)$ and discriminators with $\lambda = (\bar{\lambda}_0, \bar{\lambda}_1)$ and $\bar{\lambda}_0 > 1 > \bar{\lambda}_1$.

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

The prior beliefs $\pi_m$ 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 $\pi_m$
is associated with a particular model \( m \subseteq \{1, \ldots, J\} \) and is defined such that human decision-makers with prior \( \pi_m \) believe that only the variables in model \( m \) contain signal for predicting the outcome of interest \( Y^* \). More concretely, \( \pi_m \) is a joint distribution over the parameters \( \{ \theta^*(g, w) : (g, w) \in \{0, 1\}^{J+1}\} \) satisfying

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

(10)

for all \( g \in \{0, 1\}, w_m \in \{0, 1\}^{|m|}, w_{-m}, w'_{-m} \in \{0, 1\}^{j-|m|} \). For compactness, write \( \theta^*_{\pi_m}(g, w_m) \equiv E_{\pi_m}[\theta^*(g, w)] \), where \( w = (w_m, w_{-m}) \). There are \( 2^J \) possible models and there is a prior \( \pi_m \) associated with each model that satisfies Equation (10). The human decision-maker’s prior \( \pi_m \) can be thought of as her predictive algorithm.

We assume that at each prior \( \pi_m \), 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 \( m \). This is a “relevance” assumption and a “sufficiency” assumption.

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

\[
\theta^*_{\pi_m}(0, w_m) = \theta^*_{\pi_m}(1, w_m)
\]

for all \( w_m \in \{0, 1\}^{|m|}. \) Additionally assume that the characteristics in model \( m \) are **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

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

(11)

The market of human decision-makers is characterized by a joint distribution \( \eta(\lambda, \pi_m, C) \) over
possible combinations of preferences $\lambda$, beliefs $\pi_m$ 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 $\eta$, 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 $\lambda$, beliefs $\pi_m$ and capacity constraint $C$. She wishes to select a decision rule that maximizes her expected payoffs subject to the capacity constraint

$$\max_{t(g,w)} \sum_{(g,w) \in \{0,1\}^{i+1}} \lambda_g \theta^*_{\pi_m}(w) t(g,w) P(g,w),$$

s.t. $$\sum_{(g,w) \in \{0,1\}^{i+1}} t(g,w) P(g,w) \leq C.$$ (12)

This is 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 $1 \{ \theta^*_{\pi_m}(w) > \tau(g;C,\lambda) \}$, 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 $\lambda = (\bar{\lambda}_0, \bar{\lambda}_1)$ and $\bar{\lambda}_0 > 1 > \bar{\lambda}_1$, then she applies a higher threshold for admission to the disadvantaged group. This formalizes the sense in which such a human decision-maker is discriminatory.

4.3 The social planner’s regulation problem

The social welfare function for a given screening problem is defined as before in Equation (4), where we now assume that the social planner’s welfare weights $\psi = (\psi_0, \psi_1)$ satisfy $\psi_0 < 1 < \psi_1$. Proposition 1 implies that the social planner wishes to apply a lower threshold for admission to the disadvantaged group at her first-best decision rule, and in this sense, the social planner has an explicit equity motive.

We assume that the social planner’s preferences are aligned with the preferences of non-discriminatory human decision-makers, meaning that the preferred rank-ordering of the non-discriminatory human decision-makers is sufficiently similar to the preferred rank-ordering of the social planner.
Assumption 2 (Alignment). The social planner’s preferences are aligned with non-discriminatory human decision-makers at each prior beliefs $\pi_m$, meaning that if $w_m, w'_m$ satisfy $\theta^*_{\pi_m}(w_m) < \theta^*_{\pi_m}(w'_m)$, then

$$\psi_1 \theta^*_{\pi_m}(w_m) < \psi_0 \theta^*_{\pi_m}(w'_m).$$

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 describes a status quo in which the social planner’s equity preference is only binding relative to discriminatory human decision-makers.

The social planner does not directly observe the preferences $\lambda$, the beliefs $\pi_m$ nor the capacity constraint $C$ of any given human decision-maker. She only knows the joint distribution $\eta$ of $(\lambda, \pi_m, C)$ in the market of human decision-makers. The social planner’s payoffs are summarized by the aggregate social welfare function

$$\int_C \left( \sum_{(g,w)\in\{0,1\}^{J+1}} \psi \mathbb{E}_{\eta} \left[ \theta^*_{\pi_m}(w) t(g,w) \right] P(g,w) \right) h(C) dC. \quad (13)$$

Given that the social planner’s preferences $\psi$ do not equal the human decision-makers’ preferences $\lambda$, the optimal decision rule 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.

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 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

\[ t(g, w_m, w_m) = t(g', w_m, w_m), \]

for all \( g, g' \in \{0, 1\} \).

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.

**Remark 3.** We are crucially assuming that these model regulations are enforceable. This is a strong assumption as it 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. We discuss the strength of this assumption in more detail in Section C of the Appendix. To do so, we show that testing whether a decision rule uses group membership is equivalent to testing whether the admissions decisions are conditionally independent of group membership given the characteristics. In general, this is a difficult null hypothesis to test in finite samples, highlighting the additional challenge of detecting discrimination from realized admissions decisions.

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 \( \lambda \), beliefs \( \pi^{\hat{m}} \) and capacity constraint \( C \). At model controls \( m \), she now maximizes

\[
\sum_{g \in \{0, 1\}} \sum_{w_m \in \{0, 1\}^{\mid m \mid}} \left\{ \lambda g \sum_{w_{-m} \in \{0, 1\}^{\mid -m \mid}} \theta^{\pi^{\hat{m}}}_{m}(w_m, w_{-m}) P(g, w_m, w_{-m}) \right\} t(g, w_m) = t_{\lambda, C}^{\pi^{\hat{m}}}(g, w_m) P(g, w_m). \tag{14}
\]

The human decision-maker now rank-orders the population based upon \( \lambda g \sum_{w_m \in \{0, 1\}^{\mid m \mid}} \theta^{\pi^{\hat{m}}}_{m}(W_m, W_{-m}) \mid W_m = w_m, G = g \) as she must now pool together individuals that share the same characteristics in model \( m \). Let \( t_{\lambda, C}^{\pi^{\hat{m}}}(g, w_m) \) denote the decision rule that would be selected by a human decision-maker with preferences \( \lambda \), beliefs \( \pi^{\hat{m}} \) 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
The human decision-maker maximizes
\[
\sum_{w_m \in \{0,1\}} \left\{ \sum_{g \in \{0,1\}} \sum_{w_{-m} \in \{0,1\}} \lambda_g \theta^{\pi_{m}}_m (w_m, w_{-m}) P(w_{-m}|w_m, g) P(g|w_m) \right\} P(w_m) t(w_m) =
\]
\[
\sum_{w_m \in \{0,1\}} \left\{ \sum_{g \in \{0,1\}} \lambda_g \mathbb{E}_g \left[ \theta^{\pi_{m}}_m (W_m, W_{-m}) \mid W_m = w_m, G = g \right] P(g|w_m) \right\} t(w_m) P(w_m).
\]

The human decision-maker rank-orders the population using
\[
\lambda_g \mathbb{E}_g \left[ \theta^{\pi_{m}}_m (W_m, W_{-m}) \mid W_m = w_m, G = g \right] P(g|w_m)
\]
as she must further pool individuals across groups. Let \(t_{\tilde{m}}(w; m)\) 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 social planner’s 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 the human decision-makers as given. That is, she solves
\[
m^* = \arg \max_{m \subseteq \{1, \ldots, J\}} \int_C \left( \sum_{(g,w) \in \{0,1\}^{J+1}} \psi_g \mathbb{E}_\eta \left[ \theta^{\pi_{m}}_m (w) t_{\tilde{m}}^n (g; w; m) P(g, w) \right] \right) h(C) dC.
\]

If she additionally bans human decision-makers from using group membership, she solves
\[
m^* = \arg \max_{m \subseteq \{1, \ldots, J\}} \int_C \left( \sum_{(g,w) \in \{0,1\}^{J+1}} \psi_g \mathbb{E}_\eta \left[ \theta^{\pi_{m}}_m (w) t_{\tilde{m}}^n (g; w; m) P(g, w) \right] \right) h(C) dC.
\]

The solution \(m^*\) is the social planner’s **second-best model regulations**.

Finally, the “level of discrimination” at model controls \(m\) equals to the fraction of discriminatory human decision-makers that select a decision rule that differs from non-discriminatory human decision-makers with the same beliefs and capacity constraint.

**Definition 8.** The level of discrimination at model controls \(m\) equals
\[
\Delta(m) \equiv \mathbb{P} \left\{ t_{\tilde{m}}^n (m) \neq t_{(1,1),\tilde{m}}^n (m) \mid \lambda = (\tilde{\lambda}_0, \tilde{\lambda}_1) \right\},
\]
where \(\mathbb{P} \{ \cdot \mid \lambda = (\tilde{\lambda}_0, \tilde{\lambda}_1) \}\) is the conditional joint distribution of beliefs \(\pi_m\) 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 $\pi_m$, 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 $\theta^*_\pi_m(w) > \theta^*_\pi_m(w')$.

This is a slightly stronger assumption than the “disadvantage condition” that appears in Kleinberg and Mullainathan (2019).

How exactly the social planner selects model regulations in the second-best problem 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 $\pi_m$ across the market of human decision-makers. Therefore, in order to build intuition, we begin by first 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 $\mathcal{M} = \{m : \tilde{m} \subseteq m\}$ 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. Therefore, 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 of this result proceeds by showing 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. However, 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 $W \in \{0, 1\}^J$.

In other words, provided that the 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

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

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

Proposition 7 rewrites the aggregate social welfare function in order to illustrate the flexibility trade-off. The first term depends on the difference between the social planner’s first-best decision rule at beliefs \(\pi_{\bar{m}}\) and the decision rule that non-discriminatory human decision-makers with beliefs \(\pi_{\bar{m}}\) would select at model controls \(m\). Under the alignment assumption (Assumption 2), for model controls satisfying \(\bar{m} \subseteq m\), 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 will tend to decline towards 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 \(\pi_{\bar{m}}\) and model controls \(m\). These differ only because of the different preferences \(\lambda\) between these human decision-makers. Once the model controls are such that \(\bar{m} \subset m\), 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 and this force lowers the aggregate social welfare function. The second term therefore captures the effect that more permissive model regulations makes it easier for discriminatory human decision-makers to select decision rules that generate disparate impact.

If there were no discriminatory human decision-makers, then \(\eta(D) = 0\) and the aggregate social welfare function only depends on the first term in Proposition 7. In this case, the social planner would
find it optimal to let all human decision-makers use any characteristic, implementing no model regulations whatsoever (Proposition 4). Since the non-discriminatory human decision-makers are sufficiently aligned with the preferences of the social planner, it is optimal for the social planner to let them admit according to their preferred rank-ordering. This highlights that the presence of discriminatory human decision-makers is necessary in order for the flexibility tradeoff to be present.

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 $\pi_m$, there exists a pair of characteristics $w_m, w_m'$ such that

$$\theta_{\hat{\pi}_m}^*(w_m) > \theta_{\hat{\pi}_m}^*(w_m'), \text{ and } \bar{\lambda}(w')\theta_{\hat{\pi}_m}^*(w_m') > \bar{\lambda}(w)\theta_{\hat{\pi}_m}^*(w_m),$$

where $\bar{\lambda}(w_m) = (\lambda_0 P(0|w_m) + \lambda_1 P(1|w_m))$. 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 included in the decision loop. The social planner faced two sources of asymmetric information: over the preferences $\lambda$ and over the beliefs $\pi_m$ of the human decision-makers. Both dimensions of asymmetric information gave rise to the flexibility tradeoff as the social planner is unable to distinguish between human decision-makers that wish to use certain characteristics in their decision rule because they are believed to be predictive of the outcome of interest or because they are used to screen out members of the disadvantaged group.

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. We consider two disclosure regimes. First, we consider the case in which human
decision-makers are subject to algorithmic audits, meaning that they must disclose both their deci-
sion 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.
Moreover, the equilibrium level of discrimination is zero, meaning that the introduction of the algo-
rithm into the decision loop not only improves prediction, it simultaneously reduces discrimination
in the market. 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 receiving access to a large, randomly sampled dataset from the population of individuals
and using this training dataset to train a consistent 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 \( \mathbb{E}[Y^* | G = g, W = w] = \theta^*(g, w) \)
that satisfy

\[
\theta^*(g, w_m, w_{\neg m}) = \theta^*(g', w_m, w_{\neg m}) \\
\theta^*(g, w_m, w_{\neg m}) = \theta^*(g, w'_m, w_{\neg m})
\]

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

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
\( \theta^*(g, w_m, w_{\neg m}) \equiv \theta^*_m(w_m) \) for all \( g \in \{0, 1\}, w_{\neg m} \in \{0, 1\}^{J-|m|} \).

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

\[
\sum_{(g,w) \in \{0,1\}^{L+1}} \lambda_g \theta_m^*(w) t(g, w) P(g, w).
\]  

(16)

The human decision-maker’s optimal decision rule is a threshold rule of the form, \( \{ \theta_m^*(w) > \tau(g; C, \lambda) \} \), 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 \( \lambda \) and capacity constraints \( C \) and in a slight abuse of notation, we again denote this joint distribution by \( \eta \). 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 \((m, \lambda) \indep C \) under \( \eta \).

## 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) 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 \( \theta_m^* \) 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 particular, 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

$$m^*(m) = \arg \max_{\tilde{m} \subseteq \{1, \ldots, J\}} \int_{C} \left( \sum_{(g, w) \in \{0,1\}^{I+1}} \psi_g \mathbb{E}_{\lambda | m} \left[ \theta_{m}^{*}(w) t_{\lambda, C}(g, w; \tilde{m}) P(g, w) \right] \right) h(C) dC,$$

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.** Suppose that the disadvantage condition holds. 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. In particular, 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.** Suppose that the disadvantage conditions holds. If the social planner may conduct algorithmic audits, then the equilibrium level of discrimination goes to zero with $\Delta(m^*) = 0$. 

30
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 only must 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

\[
\min_{m \in \{1, \ldots, J\}} \sum_{\tilde{m}} \left\{ \int_C \mathbb{E} \left[ \psi(W)\theta_{\tilde{m}}^*(W) \left( t_{\tilde{m},c}(W) - t_{ND,C}(W; m) \right) \right] h(C) dC \right\} \eta(\tilde{m}) 
+ \eta(D) \sum_{\tilde{m}} \left\{ \int_C \mathbb{E} \left[ \psi(W)\theta_{\tilde{m}}^*(W) \left( t_{ND,C}(W; m) - t_{D,C}(W; m) \right) \right] h(C) dC \right\} \eta(\tilde{m}|D),
\]

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 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.

**Remark 4.** While the flexibility tradeoff is still present, it is important to note that the use of algorithms implies that the social planner does not face the challenge of inferring the human decision-maker’s decision rule from finitely many admissions decisions. For example, Kleinberg et al. (2018) note that the social planner may continually query the algorithm as many times as she would like in order to learn the underlying decision rule. While we have left this component unmodelled, this may have implications for how the level of discrimination changes after the introduction of algorithms. We briefly discuss this point in Section C of the Appendix.

## 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 from observed training data and then selected the decision rule. Crucially, the training data contained only a measured outcome which may differ from the outcome of interest. The social planner’s first-best decision rule rank-orders the population using all available information and then admits individuals according to that rank-ordering using a threshold rule with group-specific 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 in the first-best problem – equity preferences only modify the decision rule, not the prediction function.

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. Moreover, discrimination goes undetected, as the equilibrium level of discrimination is strictly positive. However, with algorithmic decision-making, the social planner may learn the true prediction function used by human decision-makers through algorithmic audits. In this case, we showed that 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, we showed that 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 provide a first step to developing an economic approach to the problem of regulating algorithms. There are several avenues for future research. As discussed in the analysis of the second-best problem, we only considered a restricted set of policy instruments that may be available to the social planner. We also abstracted away from finite-sample issues related to estimation and directly assumed that the human decision-makers and social planner learned ground truth once they accessed an algorithm. Relaxing these assumptions and analyzing the second-best problem in full generality is an important task moving forward.
References


Human Discrimination and Machine Bias
Online Appendix

A Motivating the social welfare function

In this section, we provide a sketch a simple motivation for the social welfare function given in Equation 4.

As in the main text, let $\tilde{Y}$ denote a 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 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 measured outcome $\tilde{Y}$ is

$$E_t [u_g(\tilde{Y}; T) | G = g, W = w] = t(g, w) \cdot u_g(\tilde{Y}; 1) + (1 - t(g, w)) \cdot u_g(\tilde{Y}; 0).$$

The conditional expectation here is over the admissions decision $T$ holding the measured outcome $\tilde{Y}$ fixed.

The social welfare function is defined as a weighted average of individual expected utilities under the decision rule and given by

$$\sum_{(g, w) \in \{0, 1\}^{1/1}} \psi_g \left\{ \sum_{\tilde{y} \in \{0, 1\}} \left( t(g, w) \cdot u_g(\tilde{y}; 1) + (1 - t(g, w)) \cdot u_g(\tilde{y}; 0) \right) \right\} P(g, w),$$

where $P(\tilde{y}|g, w) = P \{ \tilde{Y} = \tilde{y} | G = g, W = w \}$ for $\tilde{y} \in \{0, 1\}$ and $(\psi_0, \psi_1)$ are generalized social marginal welfare weights that vary across groups. Defining $\Delta_g(\tilde{Y}) \equiv u_g(\tilde{Y}; 1) - u_g(\tilde{Y}; 0)$, it is immediate that maximizing the social welfare function is equivalent to maximizing

$$\sum_{(g, w) \in \{0, 1\}^{1/1}} \psi_g \left\{ \sum_{\tilde{y} \in \{0, 1\}} \Delta_g(\tilde{y}) P(\tilde{y}|g, w) \right\} t(g, w)P(g, w) = \sum_{(g, w) \in \{0, 1\}^{1/1}} \psi_g E \left[ \Delta_g(\tilde{Y}) | G = g, W = w \right] t(g, w)P(g, w)$$

Therefore, without loss of generality, 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 4.

B Regularity conditions for Proposition 3

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

Recall Equation (8) in Section 3.1, which defined the likelihood function of the observed training dataset $D_N$

$$L(D_N; \eta) = \Pi_{i=1}^N \tilde{Y}_i \times \eta_0(G_i, W_i)^{1-\tilde{Y}_i} \times P(G_i, W_i), \quad I(D_N; \eta) = \log(L(D_N; \eta)).$$

Define

$$\hat{f}_N = K^{-1}_N \left( -\frac{\partial^2 I(D_N; \eta)}{\partial \eta \partial \eta'} \right) K^{-1}_N$$

and $s = \hat{f}_N^{-1/2} K_N (\hat{\eta} - \hat{\eta}_N)$. 

37
where \( \hat{\eta}_N \) is the maximum likelihood estimator of \( \eta \), 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 \( \eta \) in the population.

**Assumption 4** (Assumption 1). Assume that

1. The sequence of maximum likelihood estimators \( \hat{\eta}_N \) 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: \( \hat{J}_N \xrightarrow{d} J_0 \) and \( \hat{J}_N^{-1} \xrightarrow{d} I_0^{-1} \).

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 \( \eta^* \) given \( \tilde{\eta} \) under the prior distribution \( \pi \). We additionally make the following assumption.

**Assumption 5** (Assumption 2). 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 \( N_\delta(\tilde{\eta}_0) \).

### C Detecting discrimination in practice

In Section 4.4, we analyzed the social planner’s second best problem assuming that she may enforce her chosen model regulations on the human decision-makers’ decision rules. This assumed that the social planner directly observes the selected decision rules. In practice, this is quite a strong assumption as we only observe a finite number of admissions decisions, not the decision rule itself, meaning that the social planner faces an additional inference problem in regulating discrimination. In this section, we provide a simple extension of the model to highlight the difficulties that arise when the social planner must solve this additional inference problem.

#### C.1 Enforcing model regulations in finite samples

As emphasized by Kleinberg et al. (2018), the choices of human decision-makers are themselves a black box and a substantial portion of discrimination law centers on the challenge of simply learning the decision rule used by a human decision-maker. We apply existing results in statistics and computer science to highlight that the problem of testing for violations of model regulations is, in general, very difficult if the social planner only observes a finite number of admissions decisions of the human decision-maker, rather than the decision rule itself.

Recall that from Proposition 5, the social planner would like to ban discriminatory human decision-makers from using group membership \( G \). Suppose now that the social planner observes a random sample of \( N \) admissions decisions made by a human decision-maker. How large must the sample size \( N \) be for the social planner to accurately infer whether the human decision-maker’s decision rule \( t(g,w) \) does not use group membership? We show that this testing problem is extraordinarily difficult by establishing that any statistical test with desirable size and power properties requires access to a sample size \( N \) that grows exponentially in the number of characteristics.

More formally, suppose that the human decision-maker selects a decision rule \( t(g,w) \). The social planner observes the \( N \) randomly drawn admissions decisions that are produced by the chosen decision rule \( t(g,w) \). Each admissions decision is a random vector \( (T_i, G_i, W_i) \) with support over \( \{0,1\} \times \{0,1\} \times \{0,1\} \) and probability distribution given by

\[
\mathbb{P} \{ T_i = t, G_i = g, W_i = w \} = \begin{cases} t(g,w)p(g,w) & \text{if } t = 1, \\
(1 - t(g,w))p(g,w) & \text{if } t = 0. \end{cases}
\]
Proposition 12. The null hypothesis \( H_0 \) is equivalent to the modified null hypothesis

\[
H'_0 : T_i \perp G_i \mid W_i.
\]

The proof is simple. If the decision rule does not use group membership, then it is immediate that

\[
\mathbb{P} \{ T = t, G = g \mid W = w \} = \mathbb{P} \{ T = t \mid W = w \} \times \mathbb{P} \{ G = g \mid W = w \}
\]

for all values \( t \in \{0,1\}, g \in \{0,1\}, w \in \{0,1\}^l \). To our knowledge, we are the first to notice that the statistical problem of detecting disparate treatment in finite samples is equivalent to the problem of testing for conditional independence in discrete distributions.

In general, the problem of constructing optimal tests for properties of a discrete distribution is a challenging problem in high-dimensional settings and a full examination of this problem is beyond the scope of this paper.\(^9\) Instead, we simply show that any statistical test of whether the decision rule does not use group membership with minimally desirable properties requires an observed sample size that grows exponentially in the number of observed characteristics. This result applies Theorem 1.1 in Canonne et al. (2018) and to develop this result, we first introduce some more notation.

Let \( \mathcal{P}_{T,G|W} \) denote the set of distributions on \( \{0,1\} \times \{0,1\} \times \{0,1\}^l \) that the null hypothesis \( H'_0 \), meaning that

\[
\mathcal{P}_{T,G|W} \equiv \{ \mathbb{P} \in \Delta(\{0,1\} \times \{0,1\} \times \{0,1\}^l) : (T, G, W) \sim \mathbb{P} \text{ and } (T \perp \perp G)|W \}. \]

We define what it means for a distribution to be \( \epsilon \)-far from the null hypothesis of interest and then introduce the desired properties of the statistical test.

Definition 12. The distribution \( \mathbb{P} \in \Delta(\{0,1\} \times \{0,1\} \times \{0,1\}^l) \) is \( \epsilon \)-far from \( \mathcal{P}_{T,G|W} \) if for all \( Q \in \mathcal{P}_{T,G|W} \),

\[
d_{TV}(\mathbb{P}, Q) > \epsilon,
\]

where \( d_{TV}(\mathbb{P}, Q) \) is the total variation distance between \( \mathbb{P}, Q \). We denote this by \( d_{TV}(\mathbb{P}, \mathcal{P}_{T,G|W}) \).

Definition 13. Given a sample \( D_N \) that is drawn i.i.d. from a distribution \( \mathbb{P} \in \Delta(\{0,1\} \times \{0,1\} \times \{0,1\}^l) \), we say a statistical test \( \phi(D_N) \in \{0,1\} \) is desirable if

1. If \( \mathbb{P} \in \mathcal{P}_{T,G|W} \), then it equals one with probability at least \( 2/3 \)

\[
\mathbb{P} \{ \phi = 1 \mid \mathbb{P} \in \mathcal{P}_{T,G|W} \} \geq 2/3.
\]

2. If \( d_{TV}(\mathbb{P}, \mathcal{P}_{T,G|W}) > \epsilon \), then it equals zero with probability at least \( 2/3 \)

\[
\mathbb{P} \{ \phi = 0 \mid d_{TV}(\mathbb{P}, \mathcal{P}_{T,G|W}) > \epsilon \} \geq 2/3.
\]

In other words, we would like construct a statistical test with size less than \( 1/3 \) and power against alternatives that are local to the null hypothesis in this \( \epsilon \)-far sense at least \( 2/3 \). These are quite weak requirements. For example, in classical hypothesis testing problems, we typically desire tests with size that is less than or equal to \( 0.05 \).

\(^9\)For example, such testing problems are the focus of a new, highly active literature in statistics, computer science and machine learning. See, for example, Chan et al. (2014); Diakonikolas and Kane (2016); Balakrishnan and Wasserman (2017a,b).
Proposition 13 (Theorem 1.1. in Canonne et al. (2018)). Any statistical test \( \phi(D_N) \in \{0,1\} \) of the null hypothesis \( H_0 \) with properties described in Definition 13 requires access to a sample whose size \( N \) satisfies

\[
N \propto \max \left\{ \frac{2^{1/2}}{\epsilon^2}, \min \left\{ \frac{2^{7/8}}{\epsilon^2}, \frac{2^{6/7}}{\epsilon^{8/7}} \right\} \right\}
\]

for \( \epsilon > 0 \).

Proposition 13 states that any statistical test of whether the decision rule \( t(g,w) \) uses group membership with the minimally desirable properties in Definition 13 must have access to a sample of observed admissions decisions that grows at a rate that is exponential in the number of observed characteristics \( W \). This result highlights the difficulty of enforcing model regulations if she only observes a finite number of admissions decisions. Unless the social planner observes a large number of admissions decisions, she cannot accurately detect violations of her model regulations.

For reasonable choices of \( J, \epsilon \), the bound in Proposition 13 is enormous. To fix ideas, consider the number of characteristics that may be gleaned about an applicant from their resume and job application. At a minimum, employers observe an applicant’s education and entire work history along with demographic information such as age, gender and race. Even in this simple application, it is easy to see that the number of additional characteristics \( J \) may be quite large. Table 1 shows how the bound in Proposition 13 varies as \( J, \epsilon \) vary. Notice that even for low dimensional problems (e.g., \( J = 10 \)), the sample sizes required can be quite large. In practice, such large sample sizes are rarely observed in cases where we worry about discrimination.

### Table 1: Bound on sample size in Proposition 13 as \( J, \epsilon \) vary.

<table>
<thead>
<tr>
<th>( J )</th>
<th>( \epsilon = 0.01 )</th>
<th>( \epsilon = 0.05 )</th>
<th>( \epsilon = 0.10 )</th>
<th>( \epsilon = 0.50 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>56,568</td>
<td>2,262</td>
<td>565</td>
<td>414</td>
</tr>
<tr>
<td>10</td>
<td>320,000</td>
<td>12,800</td>
<td>4,305</td>
<td>840</td>
</tr>
<tr>
<td>25</td>
<td>( 3.84 \times 10^8 )</td>
<td>( 7.69 \times 10^7 )</td>
<td>( 3.84 \times 10^7 )</td>
<td>( 6.23 \times 10^6 )</td>
</tr>
<tr>
<td>50</td>
<td>( 1.47 \times 10^{15} )</td>
<td>( 2.44 \times 10^{14} )</td>
<td>( 1.10 \times 10^{14} )</td>
<td>( 1.75 \times 10^{13} )</td>
</tr>
<tr>
<td>100</td>
<td>( 1.22 \times 10^{28} )</td>
<td>( 1.94 \times 10^{27} )</td>
<td>( 8.81 \times 10^{26} )</td>
<td>( 1.40 \times 10^{26} )</td>
</tr>
</tbody>
</table>

C.2 The level of discrimination in finite samples

The problem of inferring decision rules from a finite number of admissions decisions has important implications for the level of discrimination (Definition 8). In particular, the level of discrimination may be higher if the social planner must infer decision rules from finitely many admissions decisions relative to the case in which the social planner observes decision rules directly. We provide a simple argument to illustrate this idea.

Suppose that the market of human decision-makers only consists of discriminatory and non-discriminatory human decision-makers with beliefs \( \pi_{\{1,...,J\}} \) or ground-truth model \( \{1,...,J\} \) with the same capacity constraint. In this case, the equilibrium level of discrimination \( \Delta(m^*) \) is equal to the fraction of discriminatory human decision-makers that use group membership in their decision rules. The social planner cannot directly observe decision rules and instead observes a finite number of admissions decisions, which she uses to learn about the human decision-makers’ chosen decision rules. She wishes to test whether the observed admissions decisions were generated by a non-discriminatory human decision-maker or discriminatory decision-maker in order to target an auditing procedure through which she learns the decision rule.

Testing whether the observed admissions decisions were generated by a non-discriminatory human decision-maker or discriminatory decision-maker is equivalent to testing

\[
H_0 : t(g,w) = t_{\{1,...,J\}}^{(1,1),\mathcal{C}}(w), \quad H_1 : t(g,w) = t_{\lambda,\mathcal{C}}^{\{1,...,J\}}(g,w)
\]
If the social planner observes \( N \) i.i.d. draws of \((T_i, G_i, W_i)\), then the most powerful test at size \( \alpha \) is the likelihood ratio test by the Neyman-Pearson Lemma. That is, the social planner rejects the null hypothesis if

\[
\frac{L_N(t_{\{1,...,J\}}^{1,1,C})}{L_N(t_{\{1,...,J\}}^{1,1,L})} > c
\]

for some critical value \( c \). Suppose, for simplicity, that the social planner audits the human decision-maker whenever this test rejects. In this set-up, the power of this test is therefore the probability that a discriminatory human decision-maker is detected and one minus the power of this test is the level of discrimination in the market.

The power of this likelihood ratio test, in general, will not equal one in finite samples, and so the level of discrimination will be strictly positive. In contrast, if the social planner directly observed the decision rules, then the level of discrimination is trivially zero. Of course, this is highly simplified setting but it illustrates that, in general, the level of discrimination may be higher if the social planner must infer decision rules from finitely many admissions decisions. While this is far from a formal argument, it suggests that the introduction of algorithmic decision-making may lower the level of discrimination relative to the case of only human decision-makers. We leave further development of this intuition as an avenue for future research.

### 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

\[
\max_{t(g, w)} \sum_{(g, w)} \psi \mathbb{E}_{\pi|D_N} [\theta^*(g, w)] t(g, w) P(g, w)
\]

\[
\text{s.t. } \sum_{(g, w)} t(g, w) P(g, w) \leq C.
\]

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 \( \psi_{\pi|D_N} \mathbb{E}_{\pi|D_N} [\theta^*(g, w)] \).

Let \((g_1, w_1), \ldots, (g_N, w_M)\) denote such an ordering with \( M = 2^{l+1} \), where

\[
\psi_{\pi|D_N} \mathbb{E}_{\pi|D_N} [\theta^*_j] = \psi_{\pi|D_N} \mathbb{E}_{\pi|D_N} [\theta^*_j(g_j, w_j)]
\]

is the \( j \)-the element of the ordering and

\[
\psi_{\pi|D_N} [\theta^*_1] \geq \psi_{\pi|D_N} [\theta^*_2] \geq \ldots \geq \psi_{\pi|D_N} [\theta^*_M].
\]

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 \( \pi|D_N \). In this case, define \( \tau(C) = \psi_{\pi|D_N} \mathbb{E}_{\pi|D_N} [\theta^*_j(C)] \).
and the social planner’s optimal admissions rule can be written as

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

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

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

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 \( \tau^*(C) = \psi_j(C) + 1 \mathbb{E}_{\pi|D_N} [\theta^*_{j(C)} + 1] \).

The social planner’s optimal admissions rule can again be rewritten as

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

where the case of ties with \( \mathbb{E}_{\pi|D_N} [\theta^*(g, w)] = \frac{\tau(C)}{\psi_S} \) is handled 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 \( \eta^* \) 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 \( \eta^* \) and \( \mathcal{L}(D_N; \eta^*) \) is the likelihood conditional on \( \eta^* \), 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

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

If \( \tilde{\eta} \perp \eta \), 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 \( \hat{\eta}^{MLE} \) denote the maximum likelihood estimate of \( \hat{\eta}(g, w) \) and let \( \hat{\eta}_0 \) denote the true value of \( \hat{\eta} \) in the population. Applying the triangle inequality, we have that

\[
\| \pi(\eta^*|D_N) - \pi(\eta^*|\hat{f}_N) \|_{TV} = \| \pi(\eta^*|D_N) - \pi(\eta^*|\hat{\eta}^{MLE}_N) \|_{TV} + \| \pi(\eta^*|\hat{\eta}^{MLE}_N) - \pi(\eta^*|\hat{f}_N) \|_{TV} \\
\leq \| \pi(\eta^*|D_N) - \pi(\eta^*|\hat{\eta}^{MLE}_N(w, g)) \|_{TV} + \| \pi(\eta^*|\hat{\eta}^{MLE}_N) - \pi(\eta^*|\hat{f}_N) \|_{TV},
\]

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

\[
\| \pi(\eta^*|\hat{\eta}^{MLE}_N) - \pi(\eta^*|\hat{f}_N) \|_{TV} \to 0.
\]

To do so, define the sequence of events \( A_n = \{ ||\hat{\eta}^{MLE}_N - \hat{\eta}_0|| < \delta, ||\hat{f}_N - \hat{\eta}_0|| < \delta \} \). 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

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

where \( ||\hat{\eta}^{MLE}_N - \hat{f}_N|| \to 0 \) because again, both are consistent. Therefore, we conclude

\[
\| \pi(\eta^*|\hat{\eta}^{MLE}_N) - \pi(\eta^*|\hat{f}_N) \|_{TV} \to 0,
\]

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 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

\[
\sum_{g \in \{0,1\}} \sum_{w \in \{0,1\}^{[m]}} \left\{ \sum_{w-m \in \{0,1\}^{[-[m]]}} \theta^*_\pi_m(w, w-m) P(g, w_m, w-m) t(g, w_m) \right\},
\]

where \( P(g, w_m, w_{m-m}) = \sum_{\bar{w} \in \{0,1\}^{[-[m]]}} P(g, w_{\bar{m}}, w_{m-\bar{m}}, w_{m-m}) \). Therefore, the human decision-maker divides the population into groups divided by the characteristics \( W_m \), \( G \) and rank orders the groups using \( \theta^*_\pi_m(w_{\bar{m}}) \). Any groups with the same value of \( w_{\bar{m}} \) 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
\[
\sum_{w_m \in \{0,1\}^{|m|}} \left\{ \sum_{w_{m-m} \in \{0,1\}^{|m|-|\tilde{m}|}} \theta^*_{\pi_0}(w_{\tilde{m}}) t(w_{\tilde{m}}, w_{m-m}) P(w_{\tilde{m}}, w_{m-m}) \right\},
\]
where \(P(w_{\tilde{m}}, w_{m-m}) = P(0, w_{\tilde{m}}, w_{m-m}) + P(1, w_{\tilde{m}}, w_{m-m})\). Again, the human decision-maker divides the population into groups based upon the characteristics \(W_m\) and rank orders the groups using \(\theta^*_{\pi_0}(w_{\tilde{m}})\). Because \(\theta^*_{\pi_0}(w_{\tilde{m}})\) 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 \(m\).

Let \(m^c = \{1, \ldots, J\} - m\) be the variables outside of \(m\) and let \(m'' = m^c \cap \tilde{m}\) be the variables outside of \(m\) and within \(\tilde{m}\). At model \(m \subset \tilde{m}\), the non-discriminatory human decision-maker selects an admissions rule to maximize
\[
\sum_{g} \sum_{w_m} \left\{ \sum_{w_{m''}} \theta^*_{\pi_0}(w_m, w_{m''}) P(g, w_m, w_{m''}) \right\} t(g, w_m)
\]
\[
= \sum_{g} \sum_{w_m} \left\{ \sum_{w_{m''}} \theta^*_{\pi_0}(w_m, w_{m''}) P(g, w_m, w_{m''}) \right\} t(g, w_m)
\]
\[
= \sum_{g} \sum_{w_m} \theta^*_{\pi_0}(w_m, w_{m''}) P(g, w_m, w_{m''}) t(g, w_m)
\]
\[
= \sum_{g} \sum_{w_m} \theta^*_{\pi_0}(w_m, w_{m''}) P(w_{m''} | g, w_m) t(g, w_m) P(g, w_m)
\]
\[
= \sum_{g} \sum_{w_m} \mathbb{E} \left[ \theta^*_{\pi_0}(W_m, W_{m''}) \mid W_m = w_m, G = g \right] t(g, w_m) P(g, w_m)
\]
if she may use group membership. Therefore, she ranks according to \(\mathbb{E} \left[ \theta^*_{\pi_0}(W_m, W_{m''}) \mid W_m = w_m, G = g \right]\).
If she may not use group-membership, then she ranks according to $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ as

$$
\sum_{w_{m}} \left\{ \sum_{g} \sum_{w_{m'}} \theta_{\pi_{0}}^{*}(w_{m}, w_{m'}) P(g, w_{m}, w_{m'}) \right\} t(w_{m})
= \sum_{w_{m}} \left\{ \sum_{g} \left( \theta_{\pi_{0}}^{*}(w_{m}, w_{m'}) \sum_{g} \sum_{w_{m'}} P(g, w_{m}, w_{m'}, w_{m'-m''}) \right) \right\} t(w_{m})
= \sum_{w_{m}} \left\{ \sum_{g} \theta_{\pi_{0}}^{*}(w_{m}, w_{m'}) P(w_{m}, w_{m'}) \right\} t(w_{m})
= \sum_{w_{m}} \left\{ \sum_{g} \theta_{\pi_{0}}^{*}(w_{m}, w_{m'}) P(w_{m'} | w_{m}) \right\} t(w_{m}) P(w_{m})
= \sum_{w_{m}} E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right] t(w_{m}) P(w_{m}).
$$

If the human decision-maker may use group membership at model $m$, then the pairs $(g, w_{m'})$ are mis-ranked must satisfy one of six possibilities:

1. $\theta_{\pi_{0}}^{*}(w_{m}) = \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) > \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m}, G = g \right]$ $> E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m}, G = g' \right]$.

2. $\theta_{\pi_{0}}^{*}(w_{m}) = \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) < \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m}, G = g \right]$ $< E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m}, G = g' \right]$.

3. $\theta_{\pi_{0}}^{*}(w_{m}) < \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) = \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m}, G = g \right]$ $= E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m}, G = g' \right]$.

4. $\theta_{\pi_{0}}^{*}(w_{m}) < \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) > \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m}, G = g \right]$ $> E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m}, G = g' \right]$.

5. $\theta_{\pi_{0}}^{*}(w_{m}) > \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) = \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m}, G = g \right]$ $= E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m}, G = g' \right]$.

6. $\theta_{\pi_{0}}^{*}(w_{m}) > \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) < \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m}, G = g \right]$ $< E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m}, G = g' \right]$.

Similarly, one of these possibilities must hold if the human decision-maker may not use group membership at model $m$. They simply become

1. $\theta_{\pi_{0}}^{*}(w_{m}) = \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) > \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ $> E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m} \right]$.

2. $\theta_{\pi_{0}}^{*}(w_{m}) = \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) < \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ $< E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m} \right]$.

3. $\theta_{\pi_{0}}^{*}(w_{m}) < \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) = \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ $= E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m} \right]$.

4. $\theta_{\pi_{0}}^{*}(w_{m}) < \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) > \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ $> E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m} \right]$.

5. $\theta_{\pi_{0}}^{*}(w_{m}) > \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) = \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ $= E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m} \right]$.

6. $\theta_{\pi_{0}}^{*}(w_{m}) > \theta_{\pi_{0}}^{*}(w'_{m}), \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w_{m}) < \psi_{g}^{*} \theta_{\pi_{0}}^{*}(w'_{m})$ and $E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w_{m} \right]$ $< E \left[ \theta_{\pi_{0}}^{*}(W_{m}, W_{m'}) | W_{m} = w'_{m} \right]$.

Notice that for all cases, if $g = g'$, then the first two inequalities cannot be satisfied simultaneously. Therefore, for all of these cases, it must be that $g \neq g'$. Moreover, it must be that $g = 1, g' = 0$ for Case 1, Case 3 and Case 4. Similarly, it must be that $g = 0, g' = 1$ for Case 2, Case 5 and Case 6.

By the alignment assumption, Cases 3–6 cannot occur. For example, in case 3, $\theta_{\pi_{0}}^{*}(w_{m}) < \theta_{\pi_{0}}^{*}(w'_{m})$, then it must be that $\psi_{0} \theta_{\pi_{0}}^{*}(w_{m}) < \psi_{0} \theta_{\pi_{0}}^{*}(w'_{m})$. Similarly, in Case 5, if $\theta_{\pi_{0}}^{*}(w_{m}) > \theta_{\pi_{0}}^{*}(w'_{m})$, then it must be that $\psi_{0} \theta_{\pi_{0}}^{*}(w_{m}) > \psi_{1} \theta_{\pi_{0}}^{*}(w'_{m})$. Therefore, we focus attention on Case 1 and Case 2.
Next, by the relevance assumption, in Case 1 and Case 2, \( w_{\tilde{m}^t} = w_{\tilde{m}^t} \). Therefore, if the human decision-maker cannot use group membership at model controls \( m \), then these will be assigned the same ranking. It follows that it must be the case that the human decision-maker may use group membership.

We proceed by contradiction to show that these inequalities may not be satisfied simultaneously for Case 1 (the argument is the same for Case 2). Since \( \theta_{\pi_{\tilde{m}}}^*(w_{\tilde{m}}) = \theta_{\pi_{\tilde{m}}}^*(w_{\tilde{m}}) \), this implies that \( w_{\tilde{m}} = w_{\tilde{m}} \).

Let \( w_{\tilde{m}} = (w_m, w_{\tilde{m} - m}) \) by assumption. The disadvantage condition immediately implies that

\[
\mathbb{E} \left[ \theta_{\pi_{\tilde{m}}}^* (W_m, W_m) \right] | W_m = w_m, G = 1 < \mathbb{E} \left[ \theta_{\pi_{\tilde{m}}}^* (W_m, W_m') \right] | W_m = w_m, G = 0
\]

by Proposition 5.2 in Kleinberg and Mullainathan (2019). By contradiction, such a mis-ranking cannot occur, and so model \( m \) cannot strictly dominate model \( \tilde{m} \).

**Step 3:** 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.

Let \( \tilde{m}' = m \cap \tilde{m} \) denote the variables within both models \( m \) and \( \tilde{m} \). Let \( \tilde{m}' \) denote the variables that are within model \( \tilde{m} \) but not model \( m \). Let \( m' = m - \tilde{m}' \) denote the variables that are within model \( m \) but not model \( \tilde{m} \). Finally, let \( m'' = \{1, \ldots, J\} - (m \cup \tilde{m}) \) denote all other variables. At model \( m' \), the non-discriminatory human decision-maker selects an admissions rule to maximize

\[
\sum_g \sum_{w_m} \left\{ \sum_{w_{m'}} \sum_{w_{m''}} \theta_{\pi_{m}}^* (w_{m'}, w_{m''}) P(g, w_{m'}, w_{m''}) t(g, w_{m'}) \right\} \]

\[
= \sum_g \sum_{w_m} \sum_{w_{m'}} \sum_{w_{m''}} \left\{ \sum_{w_{m''}} \theta_{\pi_{m}}^* (w_{m'}, w_{m''}) P(g, w_{m'}, w_{m''}) t(g, w_{m'}, w_{m''}) \right\}
\]

\[
= \sum_g \sum_{w_m} \sum_{w_{m'}} \sum_{w_{m''}} \left\{ \sum_{w_{m'}} \theta_{\pi_{m}}^* (w_{m'}, w_{m''}) P(w_{m''} | g, w_{m'}, w_{m''}) t(g, w_{m'}, w_{m''}) \right\}
\]

\[
= \sum_g \sum_{w_m} \sum_{w_{m'}} \sum_{w_{m''}} \mathbb{E} \left[ \theta_{\pi_{m}}^* (w_{m'}, w_{m''}) | W_{m'} = w_{m'}, W_{m''} = w_{m''} \right] t(g, w_{m'}, w_{m''}) P(w_{m'}, w_{m''})
\]

if she may use group membership. Similarly, if she may not use group membership, then she maximizes

\[
\sum_g \sum_{w_m} \sum_{w_{m'}} \sum_{w_{m''}} \mathbb{E} \left[ \theta_{\pi_{m}}^* (w_{m'}, w_{m''}) | W_{m'} = w_{m'}, W_{m''} = w_{m''} \right] t(w_{m'}, w_{m''}) P(w_{m'}, w_{m''})
\]

The same contradiction as in Step 2 applies via an application of the disadvantage condition.

**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

\[
\sum_{w_{\tilde{m}}} \theta_{\pi_{\tilde{m}}}^* (w_{\tilde{m}}) \{\lambda_0 P(0 | w_{\tilde{m}}) + \lambda_1 P(1 | w_{\tilde{m}})\} t(w_{\tilde{m}}) P(w_{\tilde{m}}).
\]

Defining \( \lambda(w_{\tilde{m}}) = \lambda_0 P(0 | w_{\tilde{m}}) + \lambda_1 P(1 | w_{\tilde{m}}) \), the discriminatory human decision-maker ranks the population according to \( \lambda(w_{\tilde{m}}) \theta_{\pi_{\tilde{m}}}^* (w_{\tilde{m}}) \).

**Step 1:** 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,
\[ \theta^*_{P_{\tau_0}}(w_m) = \theta^*_{P_{\tau_0}}(w_{m'}) \quad \implies \quad \lambda(w_m)\theta^*_{P_{\tau_0}}(w_m) = \lambda(w_{m'})\theta^*_{P_{\tau_0}}(w_{m'}). \]
\[ \theta^*_{P_{\tau_0}}(w_m) > \theta^*_{P_{\tau_0}}(w_{m'}) \quad \implies \quad \lambda(w_m)\theta^*_{P_{\tau_0}}(w_m) > \lambda(w_{m'})\theta^*_{P_{\tau_0}}(w_{m'}). \]

First, consider the case with \( \theta^*_{P_{\tau_0}}(w_m) = \theta^*_{P_{\tau_0}}(w_{m'}) \). By the relevance assumption, \( w_m = w_{m'} \) and therefore, \( \lambda(w_m) = \lambda(w_{m'}). \) The result follows.

Second, consider the case \( \theta^*_{P_{\tau_0}}(w_m) > \theta^*_{P_{\tau_0}}(w_{m'}) \). The disadvantage condition gives that \( \frac{p(0|w_m)}{p(1|w_m)} > \frac{p(0|w_{m'})}{p(1|w_{m'})} \), and so by Bayes’ rule
\[ \frac{p(0|w_m)}{p(1|w_m)} > \frac{p(0|w_{m'})}{p(1|w_{m'})}. \]

Since \( p(0|w_m) = 1 - p(1|w_m) \), this inequality implies that
\[ p(1|w_m) < p(1|w_{m'}) \] and \( \lambda(w_m) > \lambda(w_{m'}). \)

The result follows.

**Step 2:** For any model \( m \in \mathcal{M} = \{ m : \tilde{m} \subseteq m \} \), the discriminator constructs the same rank-ordering over the population provided that she may use group membership. This argument is the same as Step 1 in the proof of Proposition 4.

**Step 3:** We show that the social planner does not prefer letting the discriminator use group status \( G \) at model control \( \tilde{m} \) to not letting her use group status.

If the social planner prefers letting the discriminatory use group status to not using group status, then there exists some pairs \( (g, w_m), (g', w_{m'}) \) such that the discriminator ranks these pairs differently than the social planner if \( g \) is banned but ranks them in accordance with the social planner’s ranking when \( g \) is not banned. That means either

**Case 1:**
\[ \lambda_g \theta^*_{P_{\tau_0}}(w_m) \geq \lambda_{g'} \theta^*_{P_{\tau_0}}(w_{m'}) \]
\[ \psi_g \theta^*_{P_{\tau_0}}(w_m) \geq \psi_{g'} \theta^*_{P_{\tau_0}}(w_{m'}) \]
\[ (\lambda_g P(g|w_m) + \lambda_{g'} P(g'|w_{m'})) \theta^*_{P_{\tau_0}}(w_m) < (\lambda_g P(g'|w_{m'}) + \lambda_{g'} P(g'|w_{m'})) \theta^*_{P_{\tau_0}}(w_{m'}) \]

**Case 2:**
\[ \lambda_g \theta^*_{P_{\tau_0}}(w_m) > \lambda_{g'} \theta^*_{P_{\tau_0}}(w_{m'}) \]
\[ \psi_g \theta^*_{P_{\tau_0}}(w_m) > \psi_{g'} \theta^*_{P_{\tau_0}}(w_{m'}) \]
\[ (\lambda_g P(g|w_m) + \lambda_{g'} P(g'|w_{m'})) \theta^*_{P_{\tau_0}}(w_m) = (\lambda_g P(g'|w_{m'}) + \lambda_{g'} P(g'|w_{m'})) \theta^*_{P_{\tau_0}}(w_{m'}) \]

Consider Case 1. First, suppose that \( g, g' \) are equal. Then, following Step 1, the disadvantage condition implies that \( \lambda_g P(g|w_m) + \lambda_{g'} P(g'|w_{m'}) \geq \lambda_{g'} P(g'|w_{m'}) + \lambda_{g'} P(g'|w_{m'}). \) This contradicts the third inequality. Second, suppose that \( g \neq g' \). If \( g = 0, g' = 1 \), it must be that
\[ 1 > \frac{\psi_0}{\psi_1} \geq \frac{\theta^*_{P_{\tau_0}}(w_{m'})}{\theta^*_{P_{\tau_0}}(w_m)} \]
Therefore, we must have that \( \theta^*_{\pi_n}(w_m) > \theta^*_{\pi_m}(w'_m) \), and the disadvantage condition implies that 
\[
\lambda_g P(g|w_m) + \lambda_{\bar{g}} P(\bar{g}|w_m) > \lambda_{g'} P(g'|w'_m) + \lambda_{\bar{g}'} P(\bar{g}'|w'_m),
\]
contradicting the third inequality. If \( g = 1, g' = 0 \), it must be that
\[
1 > \frac{\lambda_1}{\lambda_0} = \frac{\theta^*_{\pi_n}(w'_m)}{\theta^*_{\pi_m}(w_m)}.
\]

Therefore, \( \theta^*_{\pi_n}(w_m) > \theta^*_{\pi_m}(w'_m) \), and the disadvantage condition implies that 
\[
\lambda_g P(g|w_m) + \lambda_{\bar{g}} P(\bar{g}|w_m) > \lambda_{g'} P(g'|w'_m) + \lambda_{\bar{g}'} P(\bar{g}'|w'_m),
\]
contradicting the third inequality.

Consider case 2. If \( g, g' \) are equal, then again, the disadvantage condition contradicts the third inequality. So, suppose that \( g \neq g' \). If \( g = 0, g' = 1 \), then
\[
1 > \frac{\psi_0}{\psi_1} = \frac{\theta^*_{\pi_n}(w'_m)}{\theta^*_{\pi_m}(w_m)}.
\]

An application of the disadvantage condition contradicts the third inequality. Finally, the case \( g = 1, g' = 0 \) delivers a similar contradiction.

Therefore, the social planner cannot strictly prefer letting the discriminator use group status at model control \( \bar{m} \).

**Step 4:** By a similar argument to Step 1 in the proof of Proposition 9, the discriminators rank-ordering is the same at any model \( m \in \mathcal{M} = \{ m : \bar{m} \subseteq m \} \) at which she can use group membership. The argument in Step 3 then implies that the social planner cannot prefer letting the discriminator use group status at any model \( m \in \mathcal{M} \) to model control \( \bar{m} \) with group status banned.

**Step 5:** Consider a model control \( m \) that satisfies \( \bar{m} \subset m \) with group membership banned. If the social planner strictly prefers this model control to the model control \( \bar{m} \) with group membership banned, then there exists some pairs \( (g, w_m), (g', w'_m) \) such that the discriminator ranks these pairs differently than the social planner at model \( \bar{m} \) but ranks them in accordance with the social planner’s ranking at model control \( m \).

By Step 1, this means that one of the following case must hold:

1. \( \theta^*_{\pi_n}(w_m) = \theta^*_{\pi_m}(w'_m) \), \( \psi_g \theta^*_{\pi_n}(w_m) > \psi_{\bar{g}} \theta^*_{\pi_n}(w_m) \) and \( \lambda(w_m) \theta^*_{\pi_n}(w_m) > \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \)
2. \( \theta^*_{\pi_n}(w_m) = \theta^*_{\pi_m}(w'_m) \), \( \psi_{\bar{g}} \theta^*_{\pi_n}(w_m) < \psi_g \theta^*_{\pi_n}(w_m) \) and \( \lambda(w_m) \theta^*_{\pi_n}(w_m) < \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \)
3. \( \theta^*_{\pi_n}(w_m) < \theta^*_{\pi_m}(w'_m) \), \( \psi_g \theta^*_{\pi_n}(w_m) < \psi_{\bar{g}} \theta^*_{\pi_n}(w_m) \) and \( \lambda(w_m) \theta^*_{\pi_n}(w_m) = \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \).
4. \( \theta^*_{\pi_n}(w_m) < \theta^*_{\pi_m}(w'_m) \), \( \psi_{\bar{g}} \theta^*_{\pi_n}(w_m) > \psi_g \theta^*_{\pi_n}(w_m) \) and \( \lambda(w_m) \theta^*_{\pi_n}(w_m) > \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \).
5. \( \theta^*_{\pi_n}(w_m) > \theta^*_{\pi_m}(w'_m) \), \( \psi_{\bar{g}} \theta^*_{\pi_n}(w_m) = \psi_g \theta^*_{\pi_n}(w_m) \) and \( \lambda(w_m) \theta^*_{\pi_n}(w_m) = \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \).
6. \( \theta^*_{\pi_n}(w_m) > \theta^*_{\pi_m}(w'_m) \), \( \psi_g \theta^*_{\pi_n}(w_m) > \psi_{\bar{g}} \theta^*_{\pi_n}(w_m) \) and \( \lambda(w_m) \theta^*_{\pi_n}(w_m) < \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \).

By the alignment assumption, we immediately rule out Cases 3-6 as in Step 2 in the proof of Proposition 9. Focus on Case 1 as the argument for Case 2 is analogous. In Case 1, the first equality implies \( w_m = w'_m \) by the relevance assumption, and so it must that \( \lambda(w_m) \theta^*_{\pi_n}(w_m) = \lambda(w'_m) \theta^*_{\pi_n}(w'_m) \), contradicting the third inequality.

Therefore, the social planner cannot strictly prefer a model control with \( m \) with \( \bar{m} \subset m \) and group membership banned to model control \( \bar{m} \) with group membership banned.
Step 6: Consider a model control \( m \not\in \mathcal{M} = \{ m : \tilde{m} \subset m \} \). If the social planner strictly prefers model control \( m \) to model control \( \tilde{m} \), then there exists some pairs \((g, w_{m})\), \((g', w'_{\tilde{m}})\) such that the discriminator ranks these pairs differently than the social planner at model \( m \) but ranks them in accordance with the social planner’s ranking at model control \( m \).

First, consider the case where \( m \subset \tilde{m} \). Let \( m^c = \{ 1, \ldots, J \} - m \) be the variables outside of \( m \) and let \( m'' = m^c \cap \tilde{m} \) be the variables outside of \( m \) and within \( \tilde{m} \). Suppose group membership is allowed at model control \( m \). By Step 1, if group membership is allowed, then one of the following cases must hold:

1. \( \theta^*_{\pi_m}(w_m) = \theta^*_{\pi_m}(w'_m), \psi_S^*\pi_m(w_m) > \psi_S^*\pi_m(w'_m) \) and \( \mathbb{E} \left[ \lambda_S\theta^*_{\pi_m}(W_m, W_{m''}) | W_m = w_m, G = g \right] > \mathbb{E} \left[ \lambda_S\theta^*_{\pi_m}(W_m, W_{m''}) | W_m = w'_m, G = g \right] \).
2. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) = \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) < \psi_S^*\pi_{\tilde{m}}(w'_{\tilde{m}}) \) and \( \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}}, G = g \right] < \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w'_{\tilde{m}}, G = g \right] \).
3. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) < \theta^*_{\pi_m}(w_m), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) = \psi_S^*\pi_m(w_m) \) and \( \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}}, G = g \right] = \mathbb{E} \left[ \lambda_S\theta^*_{\pi_m}(W_m, W_{m''}) | W_m = w_m, G = g \right] \).
4. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) < \theta^*_{\pi_m}(w_m), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) > \psi_S^*\pi_m(w_m) \) and \( \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}}, G = g \right] > \mathbb{E} \left[ \lambda_S\theta^*_{\pi_m}(W_m, W_{m''}) | W_m = w_m, G = g \right] \).
5. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) < \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) = \psi_S^*\pi_{\tilde{m}}(w'_{\tilde{m}}) \) and \( \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}}, G = g \right] = \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w'_{\tilde{m}}, G = g \right] \).
6. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \theta^*_{\pi_{\tilde{m}}}(w'_{\tilde{m}}), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) > \psi_S^*\pi_{\tilde{m}}(w'_{\tilde{m}}) \) and \( \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}}, G = g \right] > \mathbb{E} \left[ \lambda_S\theta^*_{\pi_{\tilde{m}}}(W_{\tilde{m}}, W_{m''}) | W_{\tilde{m}} = w'_{\tilde{m}}, G = g \right] \).

The alignment condition rules out Cases 3-6 as before. Consider Case 1 as Case 2 is analogous. In Case 1, relevance implies \( w_m = w_{\tilde{m}} \) and so, \( w_m = w'_m \). This contradicts the third inequality.

Next, suppose that group membership is not allowed at model control \( m \). By Step 1, this means that one of the following cases must hold:

1. \( \theta^*_{\pi_m}(w_m) = \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}), \psi_S^*\pi_{\tilde{m}}(w_m) > \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) \) and \( \mathbb{E} \left[ \lambda(W_m, W_{m''})\theta^*_{\pi_m}(W_m, W_{m''}) | W_m = w_m \right] > \mathbb{E} \left[ \lambda(W_{\tilde{m}}, W_{m''})\theta^*_{\pi_{\tilde{m}}}(W_m, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}} \right] \).
2. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) = \theta^*_{\pi_m}(w_m), \psi_S^*\pi_m(w_{\tilde{m}}) < \psi_S^*\pi_{\tilde{m}}(w_m) \) and \( \mathbb{E} \left[ \lambda(W_m, W_{m''})\theta^*_{\pi_{\tilde{m}}}(W_m, W_{m''}) | W_m = w_m \right] < \mathbb{E} \left[ \lambda(W_{\tilde{m}}, W_{m''})\theta^*_{\pi_m}(W_m, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}} \right] \).
3. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) < \theta^*_{\pi_m}(w_m), \psi_S^*\pi_m(w_{\tilde{m}}) = \psi_S^*\pi_{\tilde{m}}(w_m) \) and \( \mathbb{E} \left[ \lambda(W_m, W_{m''})\theta^*_{\pi_{\tilde{m}}}(W_m, W_{m''}) | W_m = w_m \right] = \mathbb{E} \left[ \lambda(W_{\tilde{m}}, W_{m''})\theta^*_{\pi_m}(W_m, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}} \right] \).
4. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) < \theta^*_{\pi_m}(w_m), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) > \psi_S^*\pi_m(w_m) \) and \( \mathbb{E} \left[ \lambda(W_m, W_{m''})\theta^*_{\pi_{\tilde{m}}}(W_m, W_{m''}) | W_m = w_m \right] > \mathbb{E} \left[ \lambda(W_{\tilde{m}}, W_{m''})\theta^*_{\pi_m}(W_m, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}} \right] \).
5. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \theta^*_{\pi_m}(w_m), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) = \psi_S^*\pi_m(w_m) \) and \( \mathbb{E} \left[ \lambda(W_m, W_{m''})\theta^*_{\pi_{\tilde{m}}}(W_m, W_{m''}) | W_m = w_m \right] = \mathbb{E} \left[ \lambda(W_{\tilde{m}}, W_{m''})\theta^*_{\pi_m}(W_m, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}} \right] \).
6. \( \theta^*_{\pi_{\tilde{m}}}(w_{\tilde{m}}) > \theta^*_{\pi_m}(w_m), \psi_S^*\pi_{\tilde{m}}(w_{\tilde{m}}) > \psi_S^*\pi_m(w_m) \) and \( \mathbb{E} \left[ \lambda(W_m, W_{m''})\theta^*_{\pi_{\tilde{m}}}(W_m, W_{m''}) | W_m = w_m \right] > \mathbb{E} \left[ \lambda(W_{\tilde{m}}, W_{m''})\theta^*_{\pi_m}(W_m, W_{m''}) | W_{\tilde{m}} = w_{\tilde{m}} \right] \).

This case is analogous as the case where group membership is allowed. Finally, the case with \( m \not\subset \tilde{m} \) with \( m \cap \tilde{m} \subset \tilde{m} \) proceeds similarly. \( \square \).
**Proof of Proposition 6**

At any cutoff $C$, the discriminatory human decision-maker wishes to select an admissions rule to maximize

$$U(t; \lambda) = \sum_{w \in \{0,1\}^j} (\lambda_0 P(0|w) + \lambda_1 P(1|w)) \tilde{\theta}_h(w) t(w) P(w)$$

$$= \sum_{w \in \{0,1\}^j} \lambda(w) \tilde{\theta}_h(w) t(w) P(w),$$

where $\lambda(w) = \lambda_0 P(0|w) + \lambda_1 P(1|w)$.

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 $\lambda(w)$ varies across the population. That is, for $w = (w_m, w_{-m}), w' = (w'_m, w'_{-m})$, it may be the case that

$$\lambda(w) \tilde{\theta}_h(w) \neq \lambda(w') \tilde{\theta}_h(w'),$$

even though $\tilde{\theta}_h(w) = \tilde{\theta}_h(w')$. It is immediate that the discriminatory firm wishes to rank order the population using $\lambda_g(w) \tilde{\theta}_h(w)$, 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 $m$ and re-write it as

$$\int_C \left( \sum_{w \in \{0,1\}^j} \left[ \psi_0 P(0|w) + \psi_1 P(1|w) \right] \sum_{m} \mathbb{E}_\eta \left[ \tilde{\theta}_h(w) t^\tilde{m}_h(w; m) \right] P(w) \right) h(C) dC$$

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

where $\psi(w) = \psi_0 P(0|w) + \psi_1 P(1|w), t^\tilde{m}_{ND,C}(g; w; m)$ 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

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

where $t^\tilde{m}_{ND,C}(g, w)$ 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

\[
\int_C \left( \sum_{w \in \{0,1\}^J} \psi(w) \left\{ \sum_m \theta_m(w) \left[ t^{\eta}_{ND,C}(w; m) - t^{\eta}_{\pi,C}(w) \right] \eta(\tilde{m}, ND) P(w) \right\} \right) h(C) dC
\]

\[
+ \int_C \left( \sum_{w \in \{0,1\}^J} \psi(w) \left\{ \sum_m \theta_m(w) \left[ t^{\eta}_{D,C}(w; m) - t^{\eta}_{\pi,C}(w) \right] \right\} \eta(\tilde{m}, D) P(w) \right) h(C) dC
\]

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

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

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

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

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

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 \( C \)

\[
t^{\eta}_{\pi,C}(m^*) = t^{\eta}_{(1,1),C}(m^*).
\]

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

\[
\theta^*_{\pi,m^*}(w_{m^*}) > \theta^*_{\pi,m^*}(w_{m^*}')
\]

\[
\tilde{\lambda} \theta^*_{\pi,m^*}(w_{m^*}) > \lambda \theta^*_{\pi,m^*}(w_{m^*}).
\]
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 $\pi_{m^*}$ rank-order according to $\theta^*_{\pi_{m^*}}(w_{m^*})$ and discriminatory human decision-makers with beliefs $\pi_{m^*}$ rank-order according to $\bar{\lambda}(w_{m^*})\theta^*_{\pi_{m^*}}(w_{m^*})$. 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 $\hat{m}$ if the social planner implements model controls $m = \hat{m}$ and bans group membership.

Therefore, at the model regulations $m^*(\hat{m}) = \hat{m}$ 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 $t^m_{\bar{\lambda},C}(m^*) = t^m_{(1,1),C}(m^*)$ as all human decision-makers simply admit individuals according to their chosen rank-ordering until the capacity constraint is satisfied. □