Towards a Non-Discriminatory Algorithm in Selected Data*

David Arnold† Will Dobbie‡ Peter Hull§

March 2021

Preliminary and Incomplete: Please Do Not Circulate or Cite

Abstract

There is a growing concern that algorithmic decision-making can lead to discrimination against legally protected groups, but measuring and reducing such algorithmic discrimination is often challenging because an individual's latent qualification for treatment is often only selectively observed. We develop new quasi-experimental tools to overcome this fundamental selection challenge and both measure and reduce algorithmic discrimination in the setting of pretrial bail decisions. We first show that algorithmic discrimination can be measured and reduced using a small number of moments involving individual qualification. We then show how these moments can be estimated by extrapolating quasi-experimental variation across as-good-as-randomly assigned decision-makers. Estimates from New York City show that both a sophisticated machine learning algorithm and a simple regression-based algorithm discriminate against Black defendants, even though defendant race and ethnicity are not included in the training data. Preliminary analyses suggest algorithmic discrimination can be dramatically reduced or eliminated using our methods, with little loss of predictive accuracy.

*We thank Ashesh Rambachan and Jonathan Roth for insightful comments. Jeremy Albright, Jerray Chang, and Alexia Olaizola provided excellent research assistance. The data we analyze are provided by the New York State Division of Criminal Justice Services (DCJS), and the Office of Court Administration (OCA). The opinions, findings, and conclusions expressed in this paper are those of the authors and not those of DCJS or OCA. Neither New York State, DCJS or OCA assumes liability for its contents or use thereof. *JEL Codes*: C26, J15, K42.

†University of California, San Diego. Email: daarnold@ucsd.edu

‡Harvard Kennedy School and NBER. Email: will_dobbie@hks.harvard.edu

§University of Chicago and NBER. Email: hull@uchicago.edu

1 Introduction

Algorithms guide an increasingly large number of high-stakes decisions, including criminal risk assessment (Chohlas-Wood, 2020), resume screening (Raghavan and Barocas, 2019), and medical testing (Price II, 2019). Alongside this rise of algorithmic decision-making is a growing concern that it can entrench or worsen discrimination against legally protected groups (Angwin et al., 2016). This concern has fueled a rich theoretical literature in computer science, where algorithmic discrimination is formalized as the differential treatment of equally qualified individuals (Zafar et al., 2017; Berk et al., 2018). With algorithmic recommendations for pretrial release decisions, for example, a risk assessment tool may be racially discriminatory if it yields systematically lower risk scores for white defendants than Black defendants with equal true risk of pretrial misconduct. Algorithms that reduce or eliminate such discrimination, by imposing new fairness constraints in estimation, have also been proposed in this theoretic literature (e.g., Agarwal et al., 2018).

Bringing the theory of algorithmic discrimination to data, however, is often hampered by a difficult selection challenge. Data on an individual's latent qualification for treatment may only be available for a group of individuals who were endogenously selected for treatment by an existing human or algorithmic decision-maker. In the setting of pretrial bail decisions, this challenge arises because pretrial misconduct potential is only observed among the defendants who a judge chooses to release before trial (Lakkaraju et al., 2017; Kleinberg et al., 2018). Such selection can both introduce bias in algorithmic predictions and complicate the measurement of algorithmic discrimination, since unobserved qualification cannot be conditioned on to compare individuals of different protected classes. This selection challenge can also complicate the objective of minimizing algorithmic discrimination, as theoretical fairness constraints can be difficult to implement when qualification is not directly observed.

This paper develops new quasi-experimental tools to overcome this fundamental selection challenge and both measure and reduce discrimination in algorithmic predictions. We first show how the measurement challenge can be solved by estimating four group-specific parameters. For racial discrimination in the pretrial setting, these are the average pretrial misconduct rates of white and Black defendants and the race-specific covariances of misconduct risk and the algorithmic predictions. In Arnold, Dobbie and Hull (2020), we show how the first set of mean risk moments can be used to measure racial discrimination in individual bail judge decisions. Here, we extend this logic by showing how the additional race-specific covariances identify racial discrimination in hypothetical algorithmic release recommendations. Intuitively, these covariances capture the differential predictiveness of the algorithm within each group, which when combined with the average qualification rates and (observed) algorithmic output can be used to construct measures of algorithmic discrimination—even when individual qualification is only selectively observed.

We next show how the four key moments can be estimated by extrapolating reduced-form variation across quasi-randomly assigned decision-makers. In the pretrial setting, average pretrial misconduct risk can be extrapolated from variation in judge-specific misconduct rates for released white and Black defendants (Arnold, Dobbie and Hull, 2020). The race-specific second moments can similarly be estimated by extrapolating judge- and race-specific second moments, of misconduct and algorithmic recommendations for released white and Black defendants. We show how both sets of extrapolations can be conducted flexibly, without specifying a model of judge decision-making and without the kinds of first-stage monotonicity restrictions (e.g., Imbens and Angrist, 1994) which have received recent scrutiny in settings with quasi-random judge assignment (e.g., Frandsen, Lefgren and Leslie, 2019).

Finally, we show how similar extrapolations of quasi-experimental variation can be used to reduce or even eliminate discrimination in new algorithmic predictions that combine observational and quasi-experimental estimates. Intuitively, these new predictions invert our testing procedure by searching within the space of possible algorithms for those which maximize predictiveness while satisfying a given fairness constraint. We implement this search by a simple constrained least squares procedure which uses extrapolated estimates of first- and second-moments of individual qualification with observable characteristics. We discuss how the tradeoff between predictive efficiency and fairness can be quantified and explored with a discrimination "possibility frontier" (Agarwal et al., 2018).

We illustrate our approach to measuring and reducing algorithmic discrimination in New York City (NYC), home to one of the largest pretrial systems in the country. We find that a sophisticated machine learning algorithm, which does not train directly on defendant race or ethnicity, yields systematically lower risk scores for white defendants than Black defendants with identical pretrial misconduct potential. When calibrated to the average NYC release rate of 73 percent, the algorithm recommends an 8 percentage point (11 percent) higher release rate for white defendants than equally qualified Black defendants. This unwarranted disparity explains 77 percent of the observed racial disparity in release recommendations, grows as the algorithm becomes more lenient, and is driven by discrimination among individuals who would engage in pretrial misconduct if released. We find a similar level of algorithmic discrimination with regression-based recommendations, using a model inspired by a widely used pretrial risk assessment tool.

Preliminary analyses suggest that such algorithmic discrimination can be reduced or even eliminated with minimal loss of predictive accuracy. Correcting selection in regression-based risk score reduces their root mean squared error (RMSE) by 20 percent while also reducing unwarranted racial disparity by 3.5 percentage points. The remaining disparity of 1.5 percentage points can be eliminated with only a 2 percent increase in RMSE. While these results are preliminary, they suggest an important role for our methods in disciplining common algorithmic risk assessment tools.

This paper adds to a recent empirical literature that uses quasi-experimental variation to test for racial bias and discrimination in the criminal justice system. Arnold, Dobbie and Yang (2018) use the release tendencies of quasi-randomly assigned bail judges to test for racial bias in a conventional linear instrumental variables (IV) framework, while Marx (2018) uses a similar approach to test for racial bias at the margin of police stops. Arnold, Dobbie and Hull (2020) show how quasi-experimental judge assignment can be used to measure a more comprehensive measure of racial discrimination, which includes racial bias, statistical discrimination, and discrimination on seemingly non-race characteristics. Other recent work in this literature includes Rose (2020) and Feigenberg and Miller (2020).[1]

Our paper also adds to theoretical and empirical work on algorithmic fairness in both computer science and economics. We propose a measure of algorithmic discrimination that is closely related to the idea of "conditional procedure accuracy equality" or "equalized odds" in the computer science tradition (Zafar et al., 2017; Berk et al., 2018), and we discuss how our approach can be used to quantify alternative fairness measures such as "equality of opportunity" (Hardt, Price and Srebro, 2016), "sufficiency" (Zafar et al., 2017), and "calibration" (Kleinberg, Mullainathan and Raghavan, 2017; Obermeyer et al., 2019). An important empirical consideration in this literature is the "selective

[1]Rose (2020) shows that a policy reform that sharply reduced prison punishments for technical probation violations nearly eliminated the racial disparity in incarceration without significantly increasing the disparity in reoffending rates, suggesting that such violations are less informative predictions of risk for Black individuals on probation. Feigenberg and Miller (2020) show that Black motorists in Texas are stopped at higher rates than white motorists without any commensurate increase in contraband hit rates, suggesting that the racial disparity in search rates is inefficient.

labels problem" (Kleinberg et al., 2018; Lakkaraju et al., 2017), which may induce racial bias in algorithmic predictions. We show how this problem, which may also confound the measurement of algorithmic discrimination, can be overcome in the pretrial bail context. Our paper also relates to a recent literature on how algorithmic recommendations interact with human decision-makers; examples from the pretrial context include Stevenson and Doleac (2019) and Albright (2019).

Methodologically, this paper adds to a recent literature on estimating average treatment effects (ATEs) with multiple discrete instruments (Kowalski, 2016; Brinch, Mogstad and Wiswall, 2017; Mogstad, Santos and Torgovitsky, 2018; Hull, 2020). The average misconduct risk parameters can be seen as race-specific ATEs, of pretrial release on pretrial misconduct, with a similar interpretation for the covariance parameters. Importantly, our approach to estimating these ATEs does not impose the usual assumption of first-stage monotonicity (Imbens and Angrist, 1994; Heckman and Vytlacil, 2005), which has received recent scrutiny both in general (Mogstad, Torgovitsky and Walters, 2019) and in the specific context of judge decision-making (Mueller-Smith, 2015; Frandsen, Lefgren and Leslie, 2019; Norris, 2019). Our approach is closely related to Hull (2020), who considers non-parametric extrapolations of quasi-experimental moments in the spirit of "identification at infinity" in conventional sample selection models (Chamberlain, 1986; Heckman, 1990; Andrews and Schafgans, 1998).

The remainder of the paper is organized as follows. Section 2 presents a general empirical framework for defining, measuring, and reducing algorithmic discrimination. Section 3 summarizes our data on pretrial bail decisions and our algorithmic predictions of pretrial misconduct risk. Section 4 applies our framework and presents our findings. Section 5 concludes.

2 Empirical Framework

2.1 Setting

We begin by defining algorithmic discrimination in a general setting, with a population of individuals i differentiated by an observable group membership $R_i \in \{w, b\}$, a latent indicator variable $Y_i^* \in \{0, 1\}$, and a vector of other observables $X_i \in \mathcal{X}$. For concreteness, and in line with our empirical application, we refer to R_i as the individual's race (either white or Black) though it could equally well represent their sex, age, or membership in any other legally protected class. The latent variable Y_i^* indicates the individual's qualification for some treatment.[2] For example, in the pretrial setting, $Y_i^* = 1$ may indicate that defendant i would engage in some form of pretrial misconduct (i.e., fail to appear in court or be rearrested for a new crime) if she were released before trial (the treatment). In a medical testing context, Y_i^* may instead indicate the latent disease state of patient i. In a resume screening setting, $Y_i^* = 1$ may indicate that individual i would be appropriate for hiring by a firm if she were granted an interview.

We define an algorithm by a mapping $f(\cdot) : \mathcal{X} \to [0, 1]$ of the observable characteristics in X_i. The output of such $f(\cdot)$ may be a prediction of individual qualification Y_i^* (i.e., $f(X_i)$ may approximate $E[Y_i^* \mid X_i]$) or a recommendation for treatment based on such a prediction (i.e., $f(X_i) = \mathbf{1}[p(X_i) < \tau]$ where $p(X_i)$ approximates $E[Y_i^* \mid X_i]$ and τ is a prespecified tolerance). For example, in the pretrial setting, $f(X_i)$ may be a defendant risk score that predicts her probability of engaging in pretrial misconduct given case and defendant observables X_i. In a medical testing context, $f(X_i)$ may be

[2] We assume such qualification is binary for simplicity. We discuss extensions in Arnold, Dobbie and Hull (2020).

a treatment recommendation given the clinical information in X_i. We allow for a general mapping which nests both simple (e.g., regression-based) predictions of Y_i^* and more elaborate machine learning procedures. The algorithm's inputs X_i may either contain or exclude group membership R_i.

Building on Arnold, Dobbie and Hull (2020), we define algorithmic discrimination by the disparity in $f(X_i)$ among equally-qualified white and Black individuals:

$$\Delta = E[E[f(X_i) \mid R_i = w, Y_i^*] - E[f(X_i) \mid R_i = b, Y_i^*]] \tag{1}$$

The inner difference in Δ compares the average algorithmic output of white and Black individuals holding fixed their true qualification Y_i^*. The outer expectation in Δ averages this comparison over the marginal qualification distribution. We say there is algorithmic discrimination when $\Delta \neq 0$ and no algorithmic discrimination when $\Delta = 0$. In the pretrial context where $f(X_i)$ is an algorithmic risk score, $\Delta < 0$ means what white defendants tend to be rated as lower risk than Black defendants with equal misconduct potential. When $f(X_i)$ instead is the algorithms release recommendation, discrimination against Black defendants is captured by $\Delta > 0$. In an employment context, $\Delta \neq 0$ may mean that the algorithm recommends interviews at systematically different rates for equally qualified white and Black job applicants, while in a medical testing context $\Delta \neq 0$ may mean that the algorithm recommends treatment at systematically different rates for equally sick white and Black patients.

This definition of algorithmic discrimination relates to idea of "conditional procedure accuracy equality" or "equalized odds" in the computer science literature (Zafar et al., 2017; Berk et al., 2018). To see this, suppose the output of the algorithm is a treatment recommendation (i.e., $f(X_i) \in \{0, 1\}$) such that it solves a binary classification problem. Such an algorithm satisfies equalized odds when it yields the same true- and false-negative rates across race.[3] Taking the example of racial discrimination, if we interpret $Y_i^* = 1$ as an adverse state (as in the pretrial setting), then Δ is an weighted average of racial disparities in true-negative rates $\delta_r^T = Pr(f(X_i) = 1 \mid Y_i^* = 0, R_i = r)$ and false negative rates $\delta_r^F = Pr(f(X_i) = 1 \mid Y_i^* = 1, R_i = r)$:

$$\Delta = (\delta_w^T - \delta_b^T)(1 - \bar{\mu}) + (\delta_w^F - \delta_b^F)\bar{\mu} \tag{2}$$

where the weights are given by the average qualification rate $\bar{\mu} = E[Y_i^*]$.[4] More generally, Equation (2) holds with $\delta_r^T = E[f(X_i) \mid Y_i^* = 0, R_i = r]$ and $\delta_r^F = E[f(X_i) \mid Y_i^* = 1, R_i = r]$. Thus, when R_i indicates race, algorithmic discrimination Δ is generally a weighted average of racial disparities in the average algorithmic outputs of equally qualified white and Black individuals.

Cast in this way, our Δ measure also aligns with the proposed definition of labor market discrimination in Aigner and Cain (1977) which compares the treatment of white and Black workers with the same objective level of productivity. In the context of racial discrimination in a pretrial setting, Δ analogously compares the algorithmic treatment of white and Black defendants with the same objective potential for pretrial misconduct, while in the medical testing example Δ compares the

[3] An alternative measure of fairness, when $f(X_i) = \mathbf{1}[p(X_i) \leq \tau]$, is the conditional independence of $p(X_i)$ and R_i given Y_i^*. See, e.g., Agarwal, Dudik and Wu (2019). The approach we develop here can be extended to measure deviations from this condition, via similar extrapolations of reduced-form variation.

[4] Other notions of algorithmic fairness include the racial equality of false-negative rates only (Hardt, Price and Srebro, 2016), the racial equality of positive and negative predictive values (Zafar et al., 2017), and equal calibration of the algorithmic predictions Obermeyer et al. (2019). We show below how our framework can also be used to quantify these alternative measures. See Kleinberg, Mullainathan and Raghavan (2017) for a theoretical discussion of various inherent tradeoffs between these and other notions of fairness.

algorithmic treatment of white and Black patients with the same disease state. We discuss in Arnold, Dobbie and Hull (2020) how parameters like Δ capture a broad notion of discrimination arising from both accurate statistical discrimination (e.g., Aigner and Cain, 1977) and racially biased preferences or beliefs (e.g., Becker, 1957; Bordalo et al., 2016). We further show that $\Delta \neq 0$ can arise either when release recommendations are directly based on class membership (i.e., R_i is included in the algorithmic input X_i) or because release decisions are based on observable characteristics that are correlated with race (i.e., variables correlated with R_i are included in the algorithm's feature set X_i).[5]

Estimating algorithmic discrimination Δ is challenging because individual qualification Y_i^* is often only selectively observed. Formally, we typically observe a censored outcome $Y_i = D_i Y_i^*$, where $D_i \in \{0, 1\}$ indicates the treatment decision of an existing human or algorithmic decision-maker. In the context of bail decisions, for example, pretrial misconduct potential Y_i^* is only observed among defendants selected by a judge for release ($D_i = 1$). Individuals who are detained before trial ($D_i = 0$) cannot engage in pretrial misconduct and so $Y_i = 0$. In the medical testing setting, patients who are tested ($D_i = 1$) have their disease state revealed but untested patients do not. In an employment setting, applicants can only be deemed appropriate for hiring ($Y_i^* = 1$) if they are granted an interview ($D_i = 1$). Such non-random selection can bias algorithmic predictions of Y_i^*, for example by causing $f(X_i) = E[Y_i^* \mid X_i, D_i = 1]$ to diverge from accurate predictions $E[Y_i^* \mid X_i]$.

Kleinberg et al. (2018) refer to the endogenous observability of Y_i^* in such settings as the "selective labels problem." They consider how algorithmic predictions of qualification can be compared to human decision-making in light of this problem. We instead consider how selection complicates measurement of algorithmic discrimination and derive a new approach to overcome this challenge. Formally, non-random selection may cause a feasible measure of discrimination:

$$\Delta^S = E[E[T_i \mid R_i = w, Y_i^*, D_i = 1] - E[T_i \mid R_i = b, Y_i^*, D_i = 1] \mid D_i = 1] \tag{3}$$

to diverge from the target discrimination parameter Δ.[6] We next discuss our solution to this problem.

2.2 Identification and Estimation

Our approach to estimating algorithmic discrimination proceeds in two steps. We first show that the challenge of measuring algorithmic discrimination reduces to a challenge of estimating four group-specific moments. For racial discrimination, these moments capture the average qualification rate for each race and how qualification covaries with the algorithmic output within race. To derive this result, we use the fact that the conditional average algorithmic outputs which constitute Δ can be written:

$$\delta_r^T = \frac{E[f(X_i)Y_i^* \mid R_i = r]}{E[Y_i^* \mid R_i = r]} = \frac{\rho_r}{\mu_r} \tag{4}$$

$$\text{and} \qquad \delta_r^F = \frac{E[f(X_i)(1 - Y_i^*) \mid R_i = r]}{E[(1 - Y_i^*) \mid R_i = r]} = \frac{E[f(X_i) \mid R_i = r] - \rho_r}{1 - \mu_r} \tag{5}$$

[5] A finding of $\Delta \neq 0$ may indicate unlawful discrimination in many settings. For example, Title VII of the 1964 Civil Rights Acts prohibits employment decisions that have a disparate impact by race. In many other contexts, including bail decisions, the Equal Protection Clause of the 14th Amendment prohibits the intentional unequal treatment of equally-qualified white and Black individuals (Yang and Dobbie, 2020).

[6] Coston et al. (2020) discuss conditions for fairness metrics like Δ to be identified by selected metrics like Δ^S. These conditions are typically strong and unlikely to hold in practice. We show they do not hold empirically in our application.

where $\mu_r = E[Y_i^* \mid R_i = r]$ denotes the average qualification rate among race-r individuals and $\rho_r = E[f(X_i)Y_i^* \mid R_i = r]$ captures the race-specific second moment of the algorithmic output $f(X_i)$ and individual qualification. The weights in Δ can further be written $\bar{\mu} = \mu_w p_w + \mu_b p_b$ where $p_r = Pr(R_i = r)$. Since these racial shares and the race-specific average output $E[f(X_i) \mid R_i = r]$ are identified, these expressions show that the missing information in Δ are the four race-specific parameters $\{\mu_w, \mu_b, \rho_w, \rho_b\}$. Algorithmic discrimination can thus be measured by estimating these four parameters, without needing to measure individual qualification directly.

We next show how the key four moments (and thus Δ) can be estimated by extrapolating reduced-form variation across as-good-as-randomly assigned decision-makers, such as bail judges in the pretrial setting. Under random assignment, each judge j makes treatment decisions D_{ij} among a comparable group of individuals i of each race. We can therefore estimate a series of judge-specific misconduct rates among the defendants, of each group, that a judge releases before trial, $\tilde{\mu}_{jR_i} \equiv E[Y_i \mid D_{ij} = 1, R_i] = E[Y_i^* \mid D_{ij} = 1, R_i]$, as well as a series of judges' race-specific release rates $\pi_{jR_i} \equiv Pr(D_{ij} = 1 \mid R_i)$. We show in Arnold, Dobbie and Hull (2020) how estimates of these differentially-selected samples of race-specific average misconduct risk can be extrapolated towards judges with high release rates to estimate the average misconduct risk parameter $E[Y_i^* \mid R_i] = \mu_{R_i}$. Our insight here is that the same logic can be applied to estimate the second moments ρ_{R_i}. Instead of the released misconduct rates, we estimate and extrapolate, for each group, the judge-specific released second moments $\tilde{\rho}_{jR_i} \equiv E[f(X_i)Y_i \mid D_{ij} = 1, R_i] = E[f(X_i)Y_i^* \mid D_{ij} = 1, R_i]$ towards judges with high release rates.[7]

To build intuition for our estimation approach, it is helpful to first consider a hypothetical "supremely lenient" bail judge j^* who releases nearly all defendants assigned to her of each group. This judge's race-specific release rates are close to one, i.e., $\pi_{j^*R_i} \approx 1$, so by quasi-random assignment her race-specific released first and second moments are both close to the unselected moments: $\tilde{\mu}_{j^*R_i} \approx \mu_{R_i}$ and $\tilde{\rho}_{j^*R_i} \approx \rho_{R_i}$. The decisions of a supremely lenient and quasi-randomly assigned judge can therefore be used to estimate the four parameters that enter our discrimination measure.

In the absence of a supremely lenient judge, the required parameters can instead be extrapolated from the variation in $\tilde{\mu}_{jR_i}$ and $\tilde{\rho}_{jR_i}$ across quasi-randomly assigned judges j within groups. This approach is analogous to a standard regression discontinuity design, in which average potential outcomes are extrapolated to a treatment cutoff from nearby observations.[8] Here, selected moments are extrapolated from quasi-randomly assigned judges to the release rate cutoff of one to estimate unselected moments. Estimates may, for example, come from the vertical intercept of linear, quadratic, or local linear regressions of the selected moment estimates on estimated release rates. Crucially, as discussed in Arnold, Dobbie and Hull (2020), such extrapolation can be conducted flexibly without assuming a model of judge decision-making or imposing the strong assumption of first-stage monotonicity which is often used (and sometimes criticized) with quasi-random judge assignment.

2.3 Estimating Non-Discriminatory Risk Scores

Our measurement procedure yields an estimate of Δ for any hypothetical risk score prediction $p(X_i)$ given by a set of observables X_i. In this section, we consider how to eliminate algorithmic discrim-

[7] This second set of extrapolations is not needed to estimate discrimination in a judge's own decisions, as in Arnold, Dobbie and Hull (2020), since if $T_i = D_{ij}$ then $\rho_{R_i} = E[D_{ij}Y_i^* \mid R_i] = E[Y_i \mid D_{ij} = 1, R_i]Pr(D_{ij} \mid R_i)$ is directly estimable for each judge j.

[8] Formally, this approach draws on a classic literature on identification "at infinity" in sample selection models (Heckman, 1990; Andrews and Schafgans, 1998).

ination by selecting the most predictive $p(X_i)$ that satisfy a fairness constraint of $\Delta = 0$. We first outline this procedure before discussing possible extensions to other objectives.

We define non-discriminatory risk scores by imposing a simple fairness constraint to a conventional predictive objective. Consider a family of risk score predictions $p(X_i; \beta)$ parameterized by a vector β. For simplicity, we consider the linear class $p(X_i; \beta) = X_i'\beta$. A conventional objective for selecting β is to minimize the mean squared error (MSE) of the risk score predictions:

$$\min_{\beta} E[(Y_i^* - X_i'\beta)^2] \tag{6}$$

where the solution is given by linear regression, $\hat{\beta} = E[X_i X_i']^{-1} E[X_i Y_i^*]$. For any choice of the parameter vector β, discrimination in the resulting risk scores can be written:

$$\begin{aligned} \Delta(\beta) &= E[E[X_i'\beta \mid R_i = w, Y_i^*] - E[X_i'\beta \mid R_i = b, Y_i^*]] \\ &= \Lambda'\beta \end{aligned} \tag{7}$$

where we define $\Lambda = E[E[X_i|R_i = w, Y_i^*] - E[X_i|R_i = b, Y_i^*]]$. This vector captures disparities in the observables X_i among, say, white and Black individuals with identical qualification Y_i^*, in parallel with our discrimination measure. We then define the non-discriminatory parameter vector as:

$$\beta^* = \arg\min_{\beta} E[(Y_i^* - X_i'\beta)^2] \text{ s.t. } \Lambda'\beta = 0 \tag{8}$$

where β^* yields the optimal (MSE minimizing) non-discriminatory risk score prediction.

Our approach to calculating such non-discriminatory risk scores again proceeds in two steps. We first note that the solution to the constrained least squares problem given in Equation (8) can be written in terms of a small number of race-specific moments. Specifically, β^* satisfies the first-order conditions:

$$2E[X_i X_i']\beta^* + \lambda^*\Lambda = 2E[X_i Y_i^*] \tag{9}$$

$$\Lambda'\beta^* = 0 \tag{10}$$

where λ^* is a Lagrange multiplier. Solving this system, we obtain:

$$\beta^* = \left(I - E[X_i X_i']^{-1}\Lambda(\Lambda' E[X_i X_i']^{-1}\Lambda)^{-1}\Lambda'\right)\hat{\beta} \tag{11}$$

$$\lambda^* = 2(\Lambda' E[X_i X_i']^{-1}\Lambda)^{-1}\Lambda'\hat{\beta} \tag{12}$$

where $\hat{\beta}$ is again the solution to the unconstrained problem and I is the identity matrix. When $\Lambda = 0$, such that there are no disparities in the observables among equally qualified white and Black individuals, this expression shows that $\beta^* = \hat{\beta}$. More generally, Equation (12) shows that the optimal non-discriminatory risk score is a function of these disparities, the vector $E[X_i Y_i^*]$ capturing the covariance of observables and qualification, and the observable second moment matrix $E[X_i X_i']$. As with our primary discrimination measure Δ, the conditional disparity vector Λ is in turn a function of the mean risk parameters $E[Y_i^* \mid R_i]$ and vector of race-specific second moments $E[X_i Y_i^* \mid R_i]$, along with the observable race-specific first moments $E[X_i \mid R_i]$. Equation (12) thus shows that the challenge of estimating the optimal non-discriminatory risk score again reduces to the challenge of

estimating a set of first and second moments, here $E[Y_i^* \mid R_i]$ and $E[X_i Y_i^* \mid R_i]$.

We next note that, as with our primary discrimination measure, we can estimate the missing information in β^* (i.e., $E[Y_i^* \mid R_i]$ and $E[X_i Y_i^* \mid R_i]$) by extrapolating reduced-form variation across as-good-as-randomly assigned decision-makers. In the pretrial context, for example, we may extrapolate conditional misconduct rates across as-good-as-randomly assigned judges to estimate the mean risk parameters as discussed above. We may also extrapolate estimates of conditional second moments $E[X_i Y_i \mid D_{ij} = 1, R_i] = E[X_i Y_i^* \mid R_i]$ to estimate the unconditional second moments. Applying these estimates to the above formulas, along with the directly estimable first- and second race-specific moments of X_i and the racial shares, yields an estimate of β^*.

This approach can be used to both estimate optimal non-discriminatory risk scores and to compare the corresponding β^* to the unconstrained optimal coefficient vector $\hat{\beta}$. We note that estimating the latter alone is a contribution of our approach, since typically the regression of Y_i^* on X_i is confounded by the selective observability of qualification. By comparing $\hat{\beta}$ to β^*, one learns what individual characteristics most contribute to algorithmic discrimination relative to their predictive content. For example, in the bail context, one may find that an individual's neighborhood is somewhat predictive of her potential for pretrial misconduct but that this characteristic yields large risk score disparities conditional on misconduct potential. In this case, the optimal non-discriminatory prediction may put little weight on neighborhood differences, relative to the unconstrained prediction. In general, the cost of including each observable will be captured by the λ multipliers, which are also computable from our estimated first and second moments. We can moreover compute the overall cost of the fairness constraint, in terms of lost predictive power, by the MSE decline from $\hat{\beta}$ to β^*.

Below, we explore two extensions of this approach. First, we explore the tradeoff between fairness and predictive capacity by relaxing the strict constraint in Equation (8) to allow for a small amount of discrimination. By varying the level of Δ in the constraint one may trace out a "possibility frontier" that optimally trades off equity and efficiency considerations (Agarwal et al., 2018). Second, we show how to strengthen the fairness constraint to eliminate a stronger notion of discrimination (within the $Y_i^* = 0$ and $Y_i^* = 1$ subgroups individually, rather than on average) or to modify the constraint to accommodate alternative notions of discrimination.[9]

3 Data

We apply our quasi-experimental framework to analyze algorithmic racial discrimination in the NYC pretrial system, one of the largest in the country. Bail conditions in NYC are set by a judge at an arraignment hearing, held shortly after an arrest. Bail hearings usually last a few minutes. The judge receives detailed information on the defendant's current offense and criminal record and decides on one of several possible bail conditions. First, she can release defendants who show minimal risk on a promise to return for all court appearances, known broadly as release on recognizance (ROR). Second, she can require defendants to post some sort of bail to be released. The judge can also send higher-risk defendants to a supervised release program as an alternative to cash bail. Finally, she can detain defendants pending trial by denying bail altogether. Bail judges are granted considerable discretion in determining who should be released before trial, but they cannot discriminate against minorities and other protected classes. Judges may consider the risk that defendants will not appear

[9] Our application of this idea is ongoing and not included in this preliminary draft.

for a required court appearance (a so-called failure to appear, or FTA) or that they will engage in new criminal activity if released.

Our analysis sample is drawn from the universe of NYC arraignments made between November 1, 2008 and November 1, 2013. We describe the construction of this sample in Arnold, Dobbie and Hull (2020), where we also give more detail on the institutional background. The sample consists of 595,186 cases involving 367,434 white or Black defendants. Each case is assigned to one of 268 judges, each of whom sees at least 100 cases. We drop cases where the defendant is not charged with a felony or misdemeanor and cases that were disposed at arraignment or adjourned in contemplation of dismissal, which are likely to be dismissed by virtually every judge.[10]

Table 1 summarizes the analysis sample, both overall and by race. Panel A shows that 73.0 percent of defendants are released before trial ($D_i = 1$). The vast majority of pretrial releases are without conditions (ROR), with only 14.4 percent of defendants being released by posting an assigned bail amount. White defendants are more likely to be released than Black defendants (76.7 percent versus 69.5 percent release rate) but among released defendants, the distribution of release conditions is virtually identical. Panel B of Table 1 shows that Black defendants are 4.9 percentage points more likely to have been arrested for a new crime before trial in the past year compared to white defendants, as well as 3.0 percentage points more likely to have a prior FTA in the past year. Panel C further shows that Black defendants are 1.3 percentage points more likely to have been charged with a felony compared to white defendants, as well as 3.6 percentage points more likely to have been charged with a violent crime. Finally, Panel D shows that Black defendants who are released are 6.6 percentage points more likely to be rearrested or have an FTA than white defendants who are released (though the composition of such misconduct is similar). Importantly, and in contrast to the other statistics in Table 1, these rates of pretrial misconduct (Y_i^*) are only measured among released defendants.

Our approach exploits the quasi-random assignment of bail judges in NYC. As detailed in Arnold, Dobbie and Hull (2020), NYC uses a rotation calendar system to assign judges to arraignment shifts in each of the five county courthouses in the city, generating quasi-random variation in bail judge assignment for defendants arrested at the same time and in the same place. Appendix Table A3 verifies the conditional randomness in assignment by regressing leave-one-out estimates of judge leniency on various defendant and case characteristics, controlling for court-by-time fixed effects. Most coefficients in this balance table are small and not statistically significantly different from zero, both overall and by defendant race, and joint F-tests fail to reject the null of quasi-random assignment.

Our approach further exploits first-stage variation in judge leniency. Appendix Table A4 verifies that differential judge assignment meaningfully affects the probability an individual is released before trial, by regressing D_i on leave-one-out estimates of judge leniency and court-by-time fixed effects. A one percentage point increase in the predicted leniency of an individual's judge leads to a 0.96 percentage point increase in the probability of release, with a somewhat smaller first-stage effect for white defendants and a somewhat larger effect for Black defendants.

Our baseline analysis measures racial discrimination in algorithmic release recommendations that are based on machine learning predictions of pretrial misconduct potential. The predictions come from a gradient boosted decision tree estimated in the sample of released defendants, following Kleinberg et al. (2018). The features X_i include a number of characteristics of the current offense and prior

[10] Appendix Table A1 compares the full sample of NYC bail cases to our estimation sample. Appendix Table A2 confirms that the quasi-random judge assignment variation we exploit in estimation is not systematically related to case disposal or dismissal. Both tables are taken from Arnold, Dobbie and Hull (2020).

criminal history, but exclude certain demographic variables such as race, ethnicity, and gender. Appendix B.1 details our estimation of this model. The model yields algorithmic risk predictions $p(X_i)$ for each defendant i. We use these predictions to form release recommendations by $T_i = \mathbf{1}[p(X_i) < \tau]$ for different risk thresholds τ. Our benchmark analysis sets τ to equalize the recommended average release rate and the actual NYC release rate of 73 percent.

Appendix Figure A1 shows that the model reliably predicts pretrial misconduct potential in the sample of released defendants. We plot true misconduct risk against the risk predictions of the machine learning model across 1,000 equal-sized bins of predicted risk, along with a local linear curve of best fit. We next discuss how we evaluate racial discrimination in recommendations based on these predictions.

4 Results

4.1 Parameter Estimates

Figures 1 and 2 show our extrapolation-based estimation of the race-specific mean risk and second moment parameters, μ_r and ρ_r, for the baseline algorithmic recommendations. The horizontal axis of both figures shows estimates of judge- and race-specific release rates π_{jr}, obtained from ordinary least squares (OLS) estimates of:

$$D_i = \sum_j \alpha_j W_i Z_{ij} + \sum_j \phi_j Z_{ij} + X_i'\beta + \epsilon_i \tag{13}$$

where D_i indicates pretrial release of defendant i, $W_i = \mathbf{1}[R_i = w]$ indicates that defendant i is white, and the Z_{ij} indicate assignment of defendant i to judge j. The control vector X_i includes court-by-time fixed effects that control for the level of quasi-experimental bail judge assignment, where we demean this vector in order to include all judge indicators. The vertical axis of Figure 1 shows the corresponding estimates of judge- and race-specific misconduct rates among released defendants μ_{jr}, obtained from OLS estimates of:

$$Y_i = \sum_j \delta_j W_i Z_{ij} + \sum_j \psi_j Z_{ij} + X_i'\gamma + u_i \tag{14}$$

among released ($D_i = 1$) individuals. The vertical axis of Figure 2, Panel A, shows estimates of judge- and race-specific second moments of misconduct and risk scores among released defendants ρ_{jr}, obtained from OLS estimates of:

$$p(X_i)Y_i = \sum_j \chi_j W_i Z_{ij} + \sum_j \phi_j Z_{ij} + X_i'\zeta + v_i \tag{15}$$

again among released individuals. Finally, the vertical axis of Figure 2, Panel B, shows estimates of judge- and race-specific second moments of misconduct and release recommendations among released defendants ρ_{jr}, obtained from OLS estimates of:

$$T_i Y_i = \sum_j \omega_j W_i Z_{ij} + \sum_j \varphi_j Z_{ij} + X_i'\theta + v_i \tag{16}$$

once again among released individuals. These specifications leverage an auxiliary assumption of linear conditional expectations of D_{ij} and Y_i^* in order to tractably accommodate the conditional random assignment of bail judges in this setting.[11]

Parameter estimates come from the vertical intercept, at one, of each race-specific extrapolation of the quasi-experimental variation in these figures. We consider linear, quadratic, and local linear extrapolations and report the corresponding parameter estimates in Panels A and B of Table 2 (for risk scores $P(X_i)$) and Table 3 (for release recommendations T_i). Our mean risk estimates in both tables match those of Arnold, Dobbie and Hull (2020); in the most flexible local linear extrapolation we estimate mean risk as $\mu_w = 0.346$ and $\mu_b = 0.436$, both with standard errors of 0.016.[12] These estimates suggest that white defendants in the population are on average 0.9 percentage points less likely to engage in pretrial misconduct. The corresponding local linear estimates of the second moments are $\rho_w = 0.127$ and $\rho_b = 0.192$ in Table 2, with standard errors of 0.006 and 0.008, and $\rho_w = 0.226$ and $\rho_b = 0.213$ in Table Y with standard errors of 0.012 and 0.017 . We obtain broadly similar estimates with the linear and quadratic extrapolations. Released misconduct rates across judges and risk score second moments trend upwards with their release rates, with relatively constant slopes, while the relationship between release recommendation second moments and release rates is flatter.[13]

4.2 Algorithmic Discrimination

The third panels of Tables 2 and Table 3 report our estimates of algorithmic discrimination, in risk scores and release recommendations, respectively. These Δ estimates use the different parameter estimates in the respective Panels A and B. We obtain the discrimination estimates by applying Equations (2), (4), and (5) to the first-step parameter estimates. Our most conservative estimates in both tables come from the local linear extrapolation. For risk scores, this yields an estimate of $\Delta = -0.047$ with a standard error of 0.003 suggesting that white defendants receive on average a 4.7 percentage point lower rating than equally-risky Black defendants. For release recommendations, we estimate $\Delta = 0.079$ with a standard error of 0.007 suggesting that white defendants are on average 7.9 percentage points more likely to be recommended for release than equally-risky Black defendants. The linear and quadratic extrapolations yield a slightly higher estimates in both cases.

Figure 3 shows how our estimate of algorithmic discrimination in release recommendations varies with the risk threshold τ, and compares these estimates to the unadjusted racial disparity in release rates. The baseline estimate of 7.9, at the average release rate in NYC, is a large share (76.0 percent) of the unadjusted disparity (10.4 percentage points). The magnitude of algorithmic discrimination rises as release rates fall, remaining a roughly equal share of the unadjusted disparity. Only at thresholds that essentially release all defendants do we fail to find a significant level of discrimination.

Appendix Figures A3 and A4 show that our finding of significant discrimination in algorithmic bail decisions is not driven by the specific machine learning algorithm that predicts pretrial misconduct risk. We obtain similar estimates of the second moments ρ_w and ρ_b, and correspondingly similar

[11] If Z_i is independent of $(T_i, Y_i^*, D_{i1}, \ldots, D_{iJ}, R_i)$ given X_i and $E[T_i Y_i^* \mid D_{ij} = 1, R_i = r, X_i] = \omega_{jr} + X_i'\theta$, then $E[T_i Y_i \mid R_i, Z_i, X_i, D_i = 1]$ is linear in $(W_i Z_{i1}, \ldots, W_i Z_{iJ}, Z_{i1}, \ldots, Z_{iJ}, X_i')'$, as in Equation (16). The same logic holds for Equations (13)-(15) under analogous linearity assumptions.

[12] We obtain standard errors by a bootstrapping procedure, in which first-step estimates from Figures 1 and 2 are redrawn according to their estimated asymptotic distribution. Standard errors in second-step parameters like μ_r, ρ_r, and Δ are then given by the standard deviation the bootstrapped estimates. First-step asymptotics are robust to two-way clustering by the defendant and judge.

[13] Appendix Figure A2 shows that the local linear parameter estimates imply a stronger (more negative) covariance of true pretrial misconduct potential and algorithmic release recommendations. See Appendix B.3 for details.

(local linear) estimates of algorithmic discrimination Δ, using simpler regression-based predictions of pretrial misconduct risk. These predictions, detailed in Appendix B.2, are inspired by a widely used pretrial risk assessment tool originally developed by the Laura and John Arnold Foundation. At the baseline release rate of 73 percent, we find a 6.7 percentage point disparity in the recommended release rates of white and Black defendants with the same potential for pretrial misconduct. This discriminatory disparity is again a large share (73.6 percent) of the unadjusted release rate disparity in algorithmic recommendations (9.1 percentage points), again increases as release rates fall, and is again statistically distinguishable from zero at all but the highest release rates.

Appendix Figure A5 shows how our baseline estimates of algorithmic discrimination in release recommendations compare with naive estimates computed on the selected sample of release defendants (i.e., estimates of Equation (3)). This comparison reveals the extent of bias due to selective labels. The selected Δ^S estimates are of similar magnitude as our selection-corrected Δ estimates, and similarly increase with the algorithm's average leniency. At the baseline NYC release rate the selected estimate is lower by 1.2 percentage points, a difference that is marginally significant at conventional levels.

Finally, Appendix Figures A6 and A7 compute alternative measures of discrimination in our baseline algorithm. We first estimate the racial disparities in true- and false-negative rates that Equation (2) show are averaged together in Δ. Racial equality in false-negative rates can be seen as satisfying what is known in the computer science literature as "equality of opportunity" (Hardt, Price and Srebro, 2016), meaning that qualified white and Black defendants without pretrial misconduct potential are released at the same rate. Appendix Figure A6 shows a disparity in false-negative rates that is large and roughly constant across different release rates, while the racial disparity in true-negative rates (i.e., the release rate differential among defendants without misconduct potential) is only statistically significantly different from zero at low release rates. These results suggest that a measure of "inequality of opportunity" (that $\delta_w^T - \delta_b^T \neq 0$) could fail to detect overall racial discrimination (that $\Delta \neq 0$) in this setting. In Appendix Figure A7, we consider departures from what is known in the computer science literature as "sufficiency" (Zafar et al., 2017), meaning the racial equality of positive and negative predictive values (see Appendix B.3 for details). We estimate a negative and relatively constant degree of "insufficiency" with our baseline parameter estimates, meaning that white defendants tend to have lower pretrial misconduct rates than Black defendants conditional on the algorithm's release recommendation. These results are driven by large but noisy racial disparities among detained defendants and smaller but more precise racial disparities among released defendants.

4.3 Non-Discriminatory Risk Scores

We next consider how the large and pervaisive extent of algorithmic discrimination can be reduced or eliminated by extensions of our quasi-experimental approach. As a preliminary analysis, we consider the class of linear risk score predictions $P(X_i;\beta) = X_i'\beta$ where X_i contains the case and defendant observables from the regression model studied in Figure A4. These observables include the defendant's age at last arrest, an indicator for a prior FTA in the previous year, an indicator for a current violent offense, the interaction of current violent offense with an indictor for age exceeding 21, an indicator for a prior misdemeanor conviction, an indictor for a prior felony conviction, an indicator for any prior conviction, an indicator for a prior violent conviction, an indicator for a prior FTA at any time, an indicator for a prior incarceration, and a constant.

We first estimate race-specific second moments of the covariate vector X_i and pretrial misconduct

potential Y_i^* by local linear extrapolation of the observed variation in such second moments across the released white and Black defendants for individual judges. We adjust these judge- and race-specific moments by court and time fixed effects, following the strategy in the previous section. We then use our local linear estimates of $E[X_i Y_i^* \mid R_i]$ and observed moments of X_i to construct two risk score predictions. The first, which we call the "unselected" prediction, uses the simple linear regression $\hat{\beta} = E[X_i X_i']^{-1} E[X_i Y_i^*]$. We compare this to the "selected" prediction which can be directly estimated, $\tilde{\beta} = E[X_i X_i' \mid D_i = 1]^{-1} E[X_i Y_i \mid D_i = 1] = E[X_i X_i' \mid D_i = 1]^{-1} E[X_i Y_i^* \mid D_i = 1]$. Unlike the selected model, the unselected model minimizes mean squared error (MSE) in predicting latent misconduct potential Y_i^* from the observables in X_i. Our second "unselected + fair" risk score prediction instead minimizes MSE within the class of models with no algorithmic discrimination, using a β^* constructed from the estimated moments as in Equation (12).

Figure 4 shows the coefficients that the three models put on each observable in X_i. The most prominent contrasts are between the selected and unselected model, suggesting a non-trivial role of the fundamental selection challenge in unconstrained predictions of Y_i^*. For example, the coefficients on prior FTA last year switches from 0.21 in the selected model to 0.04 in the unselected model; we see similar reductions for the coefficients on current violent offense and prior conviction. At the same time, the unselected model puts somewhat more weight on prior FTA at any time and prior incarceration. Imposing the fairness constraint on this model does not change the lower coefficient on prior FTA last year, while increasing the magnitude of weight put on current violence offense and prior convictions.

Figure 5 shows the level of algorithmic discrimination in each of the three models, along with their root mean squared error (RMSE) in predicting misconduct potential among all defendants. We compute mean squared error (MSE), $E[(Y_i - X_i'\beta)^2]$, using the same quasi-experimental estimates of $E[X_i Y_i^* \mid R_i]$ and observed moments of X_i that we use to fit the two unselected models. We also use these estimates to plot a range of algorithmic discrimination and RMSE among the most accurate (RMSE-minimizing) models achieving each level of unwarranted disparity. This curve gives a "possibility frontier" which quantifies the tradeoff between predictive accuracy and fairness in the class of linear estimators based on X_i.

The figure shows that algorithmic discrimination in the conventional selected risk score can be significantly reduced or even eliminated with minimal loss of predictive accuracy. One way to see this is in the reduction in discrimination as we move from the selected model (3.5 percentage points) to the unselected model (2.9 percentage points), which actually *reduces* RMSE slightly. This shows that the problem of selectively observed Y_i^* plays an important role in both raising discrimination and lowering predictive accuracy. The remaining unwarranted disparity in the unselected model can furthermore be eliminated with only a minimal (2 percent) increase in RMSE.

5 Conclusion

Algorithmic discrimination is an increasingly widespread concern in many settings, but its measurement and reduction is often hampered by a fundamental selection challenge. We show that this selection challenge can be overcome with quasi-experimental variation. Specifically, we show that the challenge of measuring algorithmic racial discrimination reduces to the challenge of estimating four race-specific parameters involving algorithmic recommendations and an individual's selectively

observed qualification. We further show that these parameters can be estimated by extrapolating reduced-form variation across as-good-as-randomly assigned decision-makers, and that similar extrapolations can be used to eliminate discrimination in new algorithms. We illustrate this approach in the NYC pretrial setting, where we find large and pervasive racial discrimination in algorithmic release recommendations that do not directly use information on race. Preliminary results suggest that such discrimination can be eliminated with minimal loss of predictive accuracy.

We conclude by noting that the methods we develop to measure and reduce algorithmic discrimination in bail decisions may prove useful in several other high-stakes settings, both within and outside of the criminal justice system. One key requirement is the quasi-random assignment of decision-makers, such as judges, police officers, employers, government benefits examiners, or medical providers. A second requirement is that an individual's qualification of treatment is measurable among a subset of individuals that the decision-maker endogenously selects. Mapping these settings to the quasi-experimental approach in this paper can overcome fundamental selection challenges and bring a large theoretical literature on algorithmic (un)fairness to data.

References

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. "A Reductions Approach to Fair Classification." *Proceedings of the 35th International Conference on Machine Learning*, 60–69.

Agarwal, Alekh, Miroslav Dudik, and Zhiwei Steven Wu. 2019. "Fair Regression: Quantitative Definitions and Reduction-based Algorithms." *Proceedings of the 36th International Conference on Machine Learning*, 120–129.

Aigner, Dennis, and Glen Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, 30(2): 175–187.

Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." *Unpublished Working Paper.*

Andrews, Donald, and Marcia Schafgans. 1998. "Semiparametric Estimation of the Intercept of a Sample Selection Model." *Review of Economic Studies*, 65(3): 497–517.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica Report.*

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics*, 133(4): 1885–1932.

Arnold, David, Will Dobbie, and Peter Hull. 2020. "Measuring Racial Discrimination in Bail Decisions." *NBER Working Paper No. 26999.*

Becker, Gary S. 1957. *The Economics of Discrimination.* University of Chicago Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*, 1–42.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics*, 131(4): 1753–1794.

Brinch, Christian, Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy*, 125(4): 985–1039.

Chamberlain, Gary. 1986. "Asymptotic Efficiency in Semiparametric Models with Censoring." *Journal of Econometrics*, 32(2): 189–218.

Chohlas-Wood, Alex. 2020. "Understanding Risk Assessment Instruments in Criminal Justice." *Brookings Report.*

Coston, Amanda, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. "Counterfactual Risk Assessments, Evaluation, and Fairness." In *Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY.

Feigenberg, Benjamin, and Conrad Miller. 2020. "Racial Disparities in Motor Vehicle Searches Cannot Be Justified by Efficiency." *NBER Working Paper No. 27761.*

Frandsen, Brigham R., Lars J. Lefgren, and Emily C. Leslie. 2019. "Judging Judge Fixed Effects." *NBER Working Paper No. 25528.*

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *Proceedings of the 30th Conference on Neural Information Processing Systems*, 3323–3331.

Heckman, James J. 1990. "Varieties of Selection Bias." *American Economic Review*, 80(2): 313–318.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Hull, Peter. 2020. "Estimating Hospital Quality with Quasi-Experimental Data." *Unpublished Working Paper.*

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-Offs in Algorithmic Fairness." *Proceedings of Innovations in Theoretical Computer Science*, 1–23.

Kowalski, Amanda. 2016. "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." *NBER Working Paper No. 22363.*

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Marx, Philip. 2018. "An Absolute Test of Racial Prejudice." *Unpublished Working Paper.*

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2019. "Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions." *NBER Working Paper No. 25691.*

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference About Policy-Relevant Treatment Parameters." *Econometrica*, 86(5): 1589–1619.

Mueller-Smith, Michael. 2015. "The Criminal and Labor Market Impacts of Incarceration." *Unpublished Working Paper.*

Norris, Sam. 2019. "Examiner Inconsistency: Evidence from Refugee Appeals." *Unpublished Working Paper.*

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, 366: 447–453.

Price II, W. Nicholson. 2019. "Risks and Remedies for Artificial Intelligence in Health Care." *Brookings Report.*

Raghavan, Manish, and Solon Barocas. 2019. "Challenges for Mitigating Bias in Algorithmic Hiring." *Brookings Report.*

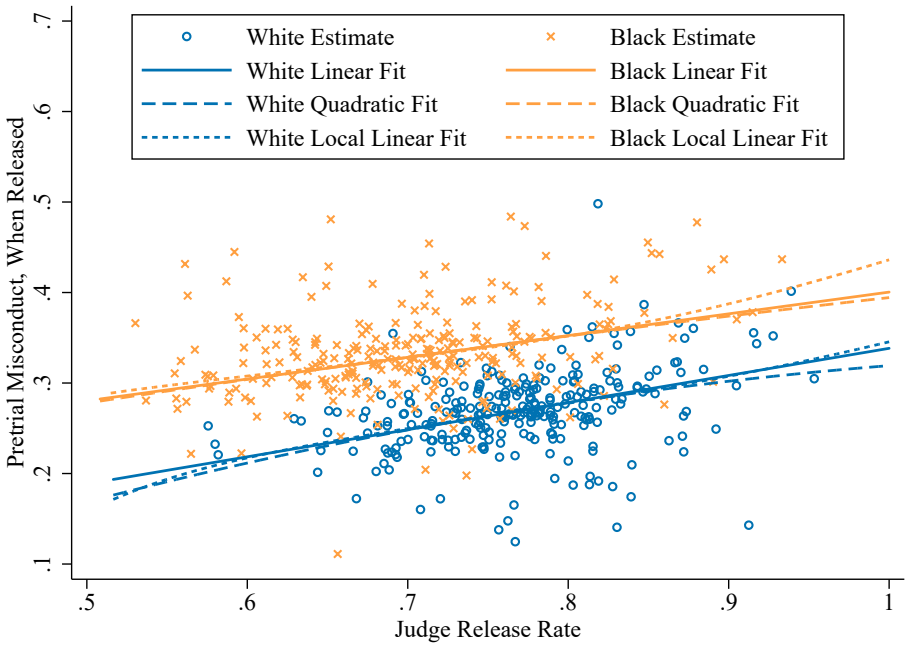
Rose, Evan. 2020. "Who Gets a Second Chance? Effectiveness and Equity in Supervision of Criminal Offenders." *Unpublished Working Paper.*

Stevenson, Megan T., and Jennifer L. Doleac. 2019. "Algorithmic Risk Assessment in the Hands of Humans." *Unpublished Working Paper.*

Yang, Crystal, and Will Dobbie. 2020. "Equal Protection Under Algorithms: A New Statistical and Legal Framework." *Michigan Law Review*, 119(1): 291–396.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna Gummadi. 2017. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment." *Proceedings of the 26th International Conference on World Wide Web.*

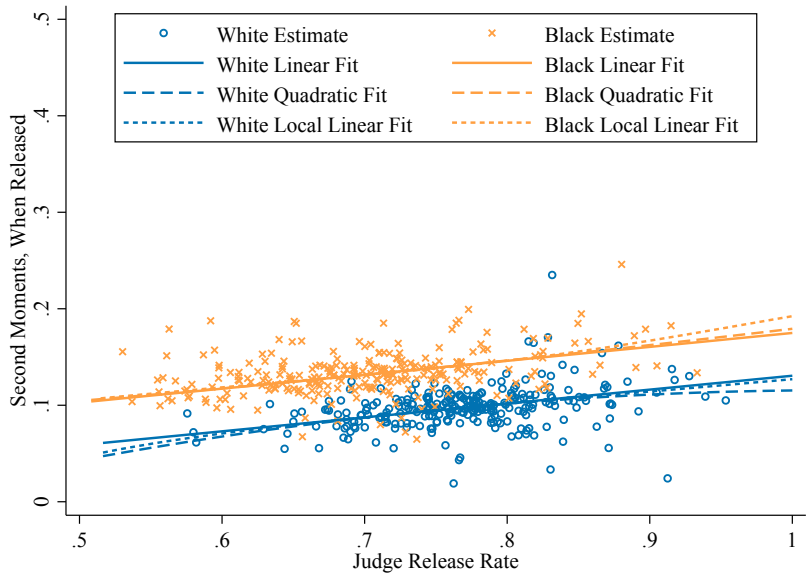
Figure 1: Extrapolating Released Misconduct Rates Across Bail Judges



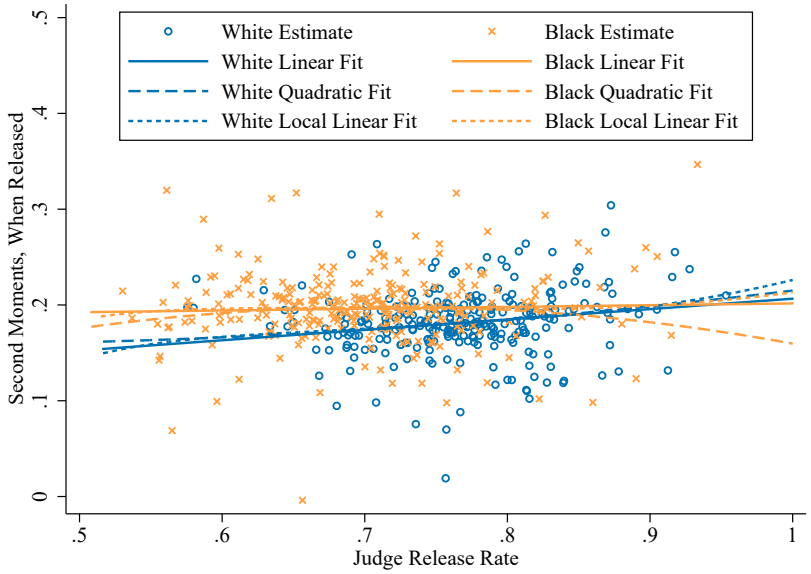
Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of pretrial misconduct among the set of released defendants. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated misconduct rate among released defendants. The local linear regressions use a Gaussian kernel with a race-specific rule-of-thumb bandwidth.

Figure 2: Extrapolating Second Moments Across Bail Judges

A. Risk Score $P(X_i)$

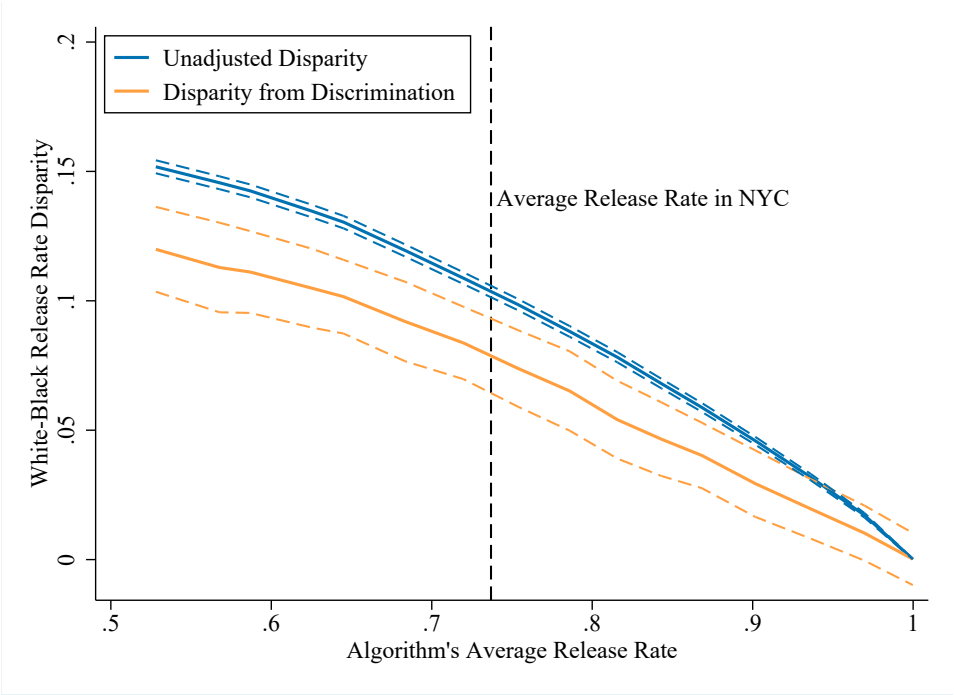


B. Release Recommendation T_i (Average NYC Release Rate)



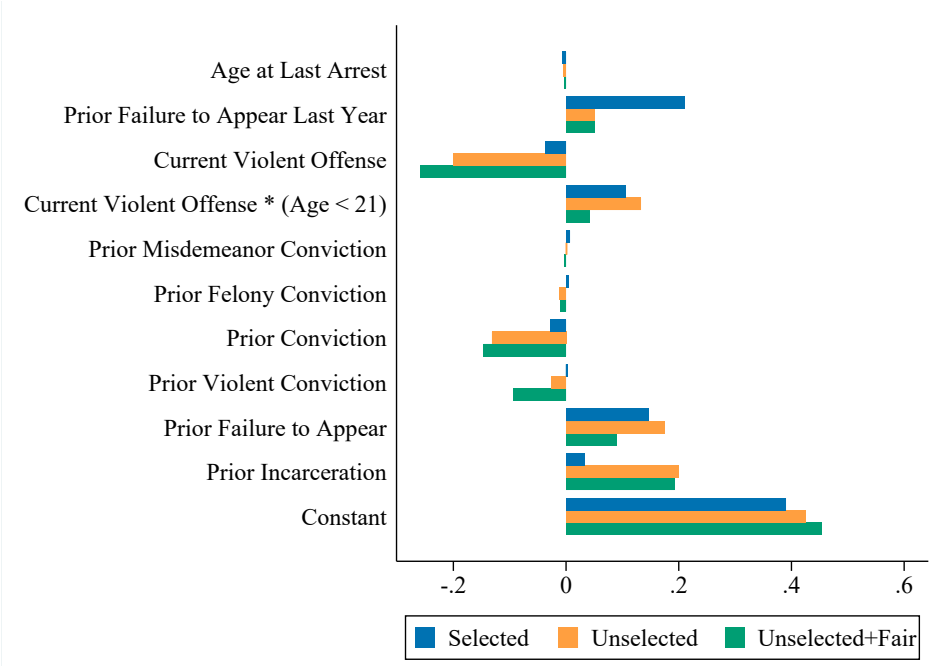
Notes. This figure plots race-specific release rates for the 268 judges in our sample against the uncentered second moments of pretrial misconduct and either risk scores (Panel A) or algorithmic recommendations (Panel B) among the set of released defendants. Algorithmic recommendations are from our baseline gradient-boosted decision tree model with a risk threshold calibrated to equalize the average recommended release rate and the average release rate in NYC. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated misconduct rate among released defendants. The local linear regressions use a Gaussian kernel with a race-specific rule-of-thumb bandwidth.

Figure 3: Discrimination in Algorithmic Release Recommendations



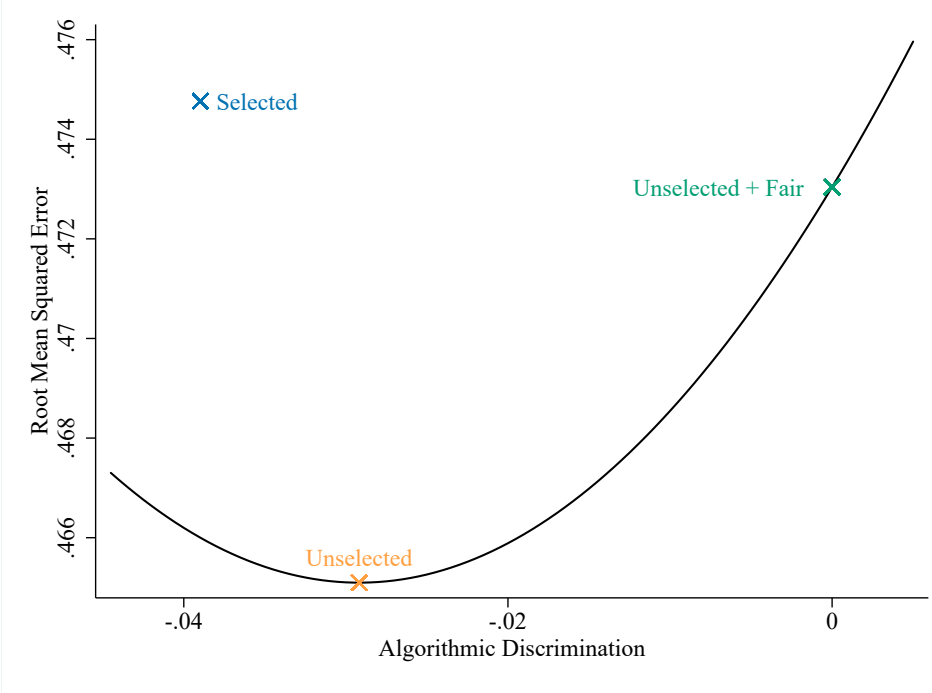
Notes. This figure plots the range of unadjusted racial disparities in algorithmic release rate recommendations, for different average release rates, along with the range of disparities due to racial discrimination. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities from discrimination are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described in the text.

Figure 4: Coefficients in Algorithmic Risk Score Predictions



Notes. This figure shows the estimated coefficients in our three regression-based risk score prediction models. The selected model is given by the regression of misconduct potential on the covariates among released defendants, and coincides with the model discussed in Appendix B.2. The unselected model uses our local linear estimates of the second moment of misconduct potential and covariates to construct the corresponding regression among all defendants. The unselected+fair model adds a linear constraint to this regression, ensuring the risk score predictions exhibit no algorithmic discrimination.

Figure 5: Accuracy vs. Discrimination in Algorithmic Risk Score Predictions



Notes. This figure plots root mean squared error of our three regression-based risk score prediction models against ther level of algorithmic discrimination. The black curve indicates the range of possible values for models that minimize root mean squared error subject to a constraint on the level of unwarranted disparity. Computations use our local linear estimates of the second moment of misconduct potential and covariates, as described in the text

Table 1: Descriptive Statistics

	All Defendants	White Defendants	Black Defendants
Panel A: Pretrial Release	(1)	(2)	(3)
Released Before Trial	0.730	0.767	0.695
Share ROR	0.852	0.852	0.851
Share Money Bail	0.144	0.144	0.145
Share Other Bail Type	0.004	0.004	0.004
Share Remanded	0.000	0.000	0.000
Panel B: Defendant Characteristics			
White	0.478	1.000	0.000
Male	0.821	0.839	0.804
Age at Arrest	31.97	32.06	31.89
Prior Rearrest	0.229	0.204	0.253
Prior FTA	0.103	0.087	0.117
Panel C: Charge Characteristics			
Number of Charges	1.150	1.184	1.118
Felony Charge	0.362	0.355	0.368
Misdemeanor Charge	0.638	0.645	0.632
Any Drug Charge	0.256	0.257	0.256
Any DUI Charge	0.046	0.067	0.027
Any Violent Charge	0.143	0.124	0.160
Any Property Charge	0.136	0.127	0.144
Panel D: Pretrial Misconduct, When Released			
Pretrial Misconduct	0.299	0.266	0.332
Share Rearrest Only	0.499	0.498	0.499
Share FTA Only	0.281	0.296	0.269
Share Rearrest and FTA	0.220	0.205	0.232
Total Cases	595,186	284,598	310,588
Cases with Defendant Released	434,201	218,256	215,945

Notes. This table summarizes the NYC analysis sample. The sample consists of bail hearings that were quasi-randomly assigned to judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records as described in the text. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. ROR (released on recognizance) is defined as being released without any conditions. FTA (failure to appear) is defined as failing to appear at a mandated court date.

Table 2: Parameter and Discrimination Estimates for Algorithmic Risk Scores $p(X_i)$

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Misconduct Risk	(1)	(2)	(3)
White Defendants	0.338	0.319	0.346
	(0.007)	(0.021)	(0.016)
Black Defendants	0.400	0.394	0.436
	(0.006)	(0.022)	(0.015)
Panel B: Risk Score Second Moment			
White Defendants	0.131	0.115	0.127
	(0.003)	(0.008)	(0.006)
Black Defendants	0.175	0.179	0.192
	(0.003)	(0.010)	(0.008)
Panel C: Disparity in Risk Scores			
Risk Score Disparity	-0.050	-0.049	-0.047
	(0.001)	(0.004)	(0.003)

Notes. Panels A and B of this table summarize estimates of race-specific mean risk and second moments of misconduct potential and the algorithmic release scores from different extrapolations of the variation in Figure 1 and Panel A of Figure 2. Panel C reports corresponding estimates of algorithmic discrimination, as defined in the text. The unadjusted gap in risk scores is equal to -0.058, indicating on average, the algorithm predicts risk is 5.8 percentage points lower for white defendants relative to Black defendants. Column 1 uses a linear extrapolation of the variation, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

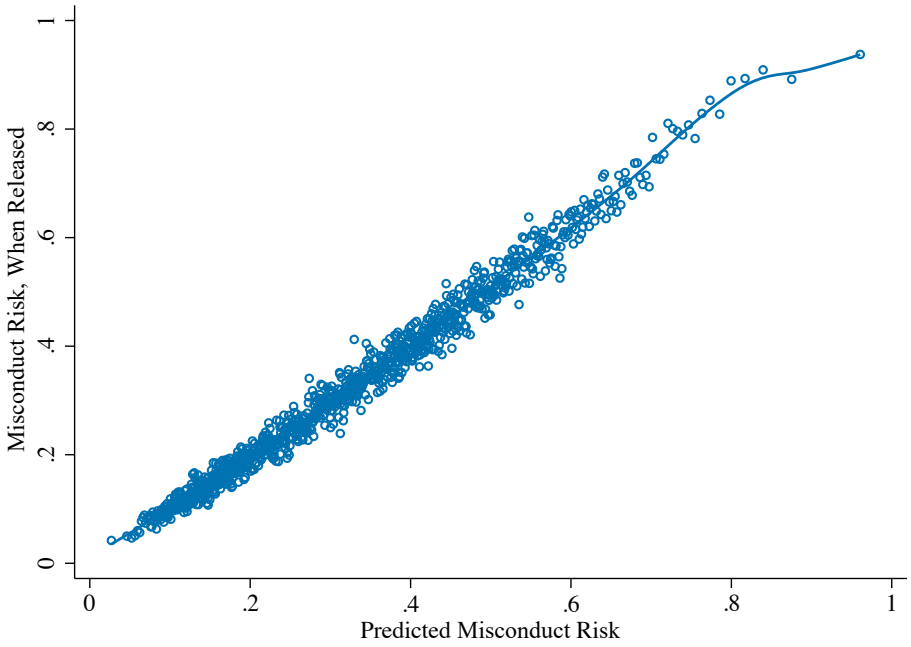
Table 3: Parameter and Discrimination Estimates for Algorithmic Recommendations T_i (Average NYC Release Rate)

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Misconduct Risk	(1)	(2)	(3)
White Defendants	0.338	0.319	0.346
	(0.007)	(0.022)	(0.016)
Black Defendants	0.400	0.394	0.436
	(0.006)	(0.020)	(0.016)
Panel B: Misconduct/Recommendation Second Moment			
White Defendants	0.207	0.215	0.226
	(0.006)	(0.019)	(0.012)
Black Defendants	0.202	0.160	0.213
	(0.006)	(0.016)	(0.017)
Panel C: Algorithmic Discrimination			
Release Rate Disparity	0.086	0.080	0.079
	(0.003)	(0.011)	(0.007)

Notes. Panels A and B of this table summarize estimates of race-specific mean risk and second moments of misconduct potential and the algorithmic release recommendation from different extrapolations of the variation in Figure 1 and Panel B of Figure 2. Panel C reports corresponding estimates of algorithmic discrimination, as defined in the text. Column 1 uses a linear extrapolation of the variation, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

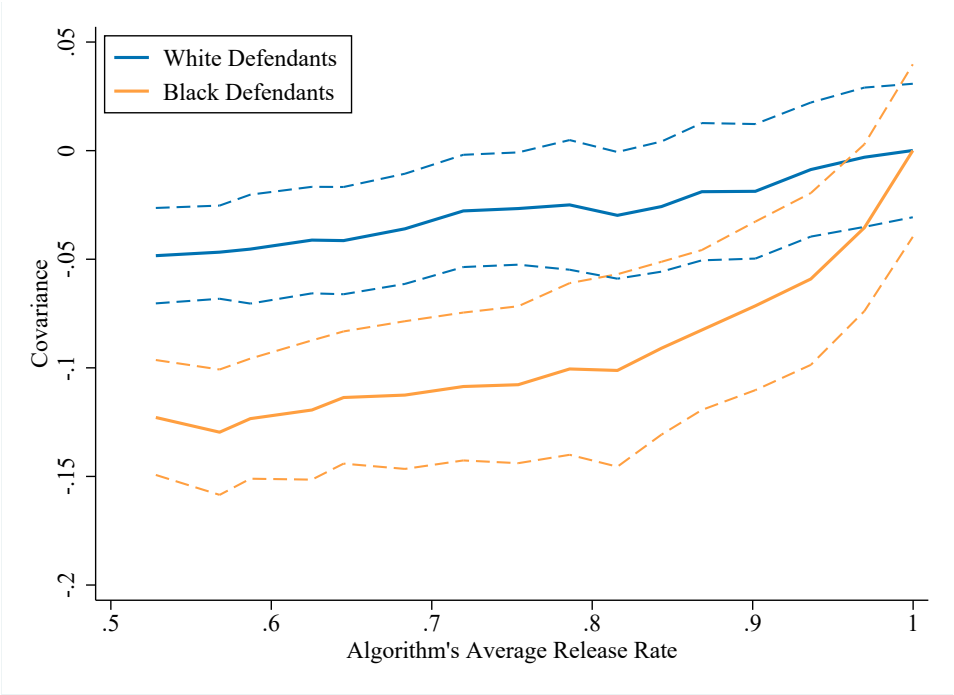
A Appendix Figures and Tables

Appendix Figure A1: Algorithmic Predictions of Pretrial Misconduct Risk



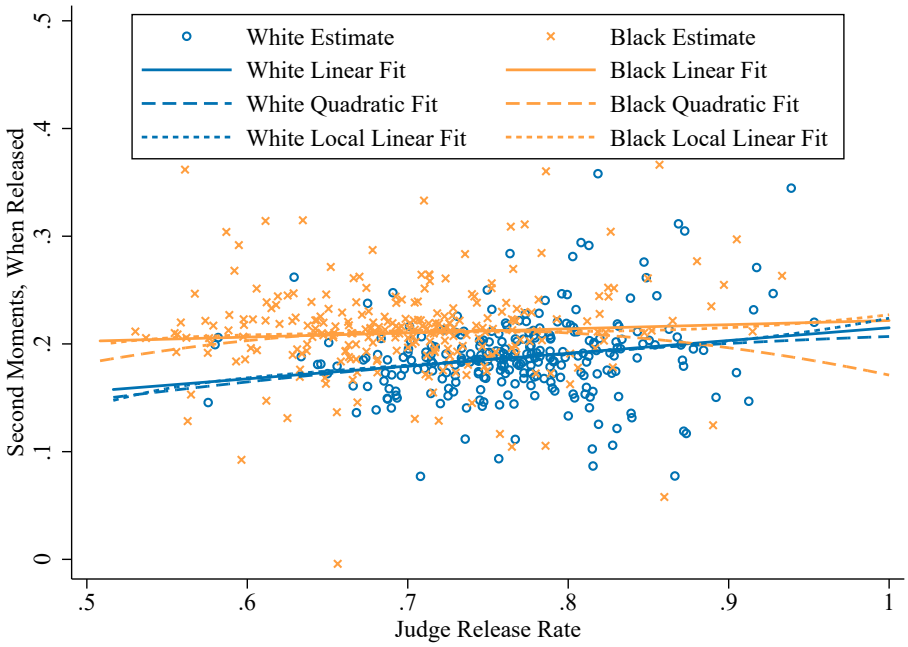
Notes. This figure plots true pretrial misconduct risk against the risk predictions of our baseline gradient-boosted decision tree algorithm among the set of released defendants in our sample. True risk is computed within 1,000 equal-sized bins of predicted risk. The curve of best fit comes from a local linear regression with a Gaussian kernel and rule-of-thumb bandwidth.

Appendix Figure A2: Covariance of Pretrial Misconduct and Algorithmic Release Recommendations



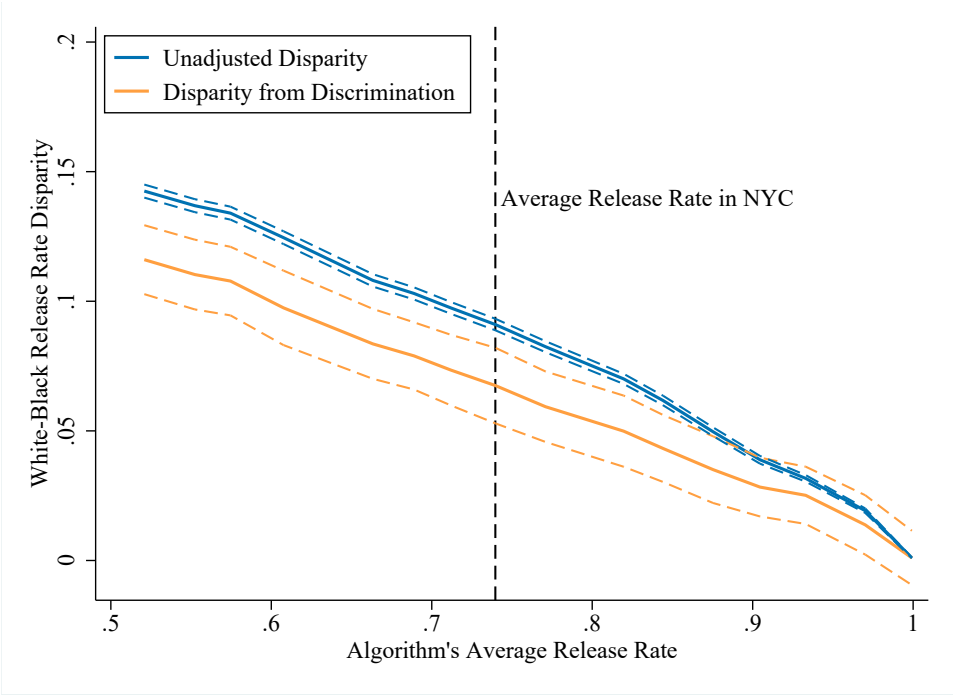
Notes. This figure plots the range of race-specific covariance between pretrial misconduct potential and algorithmic release recommendations for different average release rates. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Covariances are computed by using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described in the text.

Appendix Figure A3: Extrapolating Regression-Based Second Moments Across Bail Judges



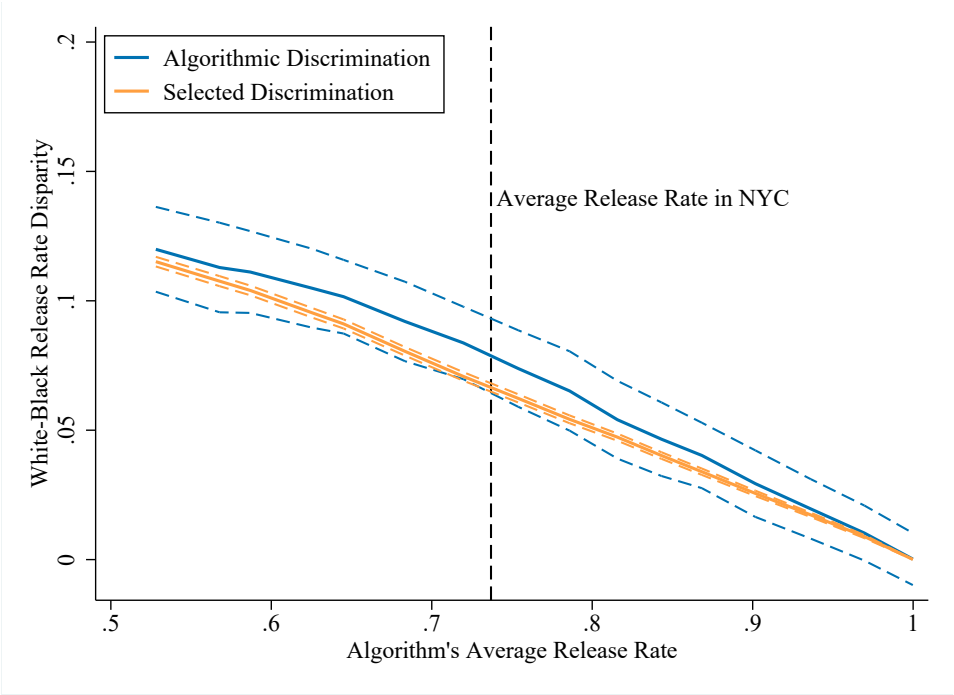
Notes. This figure plots race-specific release rates for the 268 judges in our sample against the uncentered second moments of pretrial misconduct and algorithmic recommendations among the set of released defendants. Algorithmic recommendations are from the regresison model described in the text with a risk threshold calibrated to equalize the average recommended release rate with the average release rate in NYC. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated misconduct rate among released defendants. The local linear regressions use a Gaussian kernel with a race-specific rule-of-thumb bandwidth.

Appendix Figure A4: Discrimination in Regression-Based Algorithmic Bail Decisions



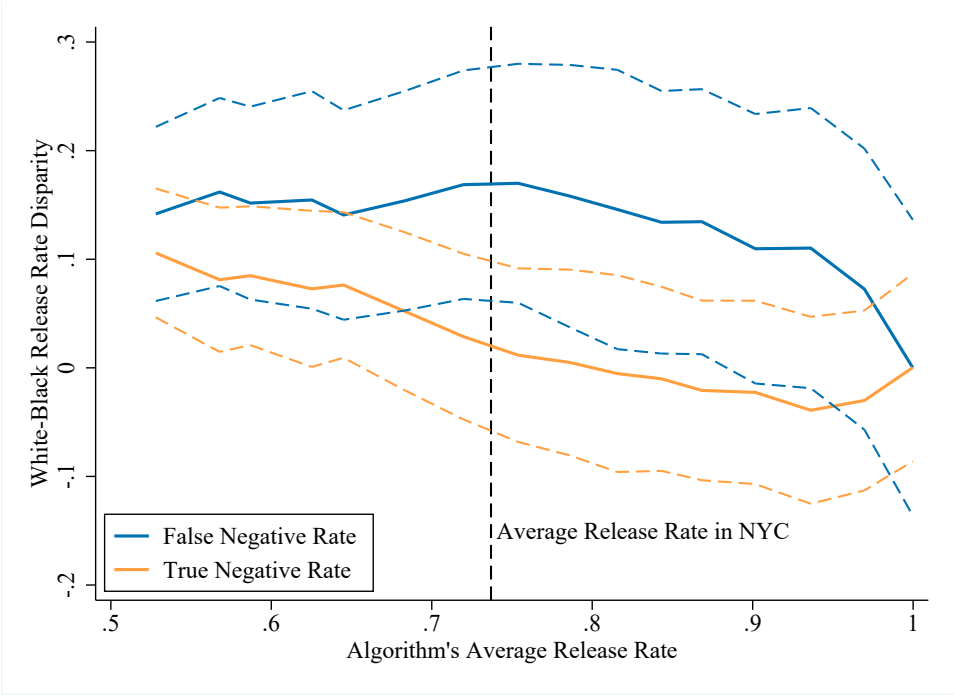
Notes. This figure plots the range of unadjusted racial disparities in algorithmic release rate recommendations, for different average release rates, along with the range of disparities due to racial discrimination. Algorithmic recommendations are from the regression model described in the text. Disparities from discrimination are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described in the text.

Appendix Figure A5: Quantifying the Bias from Selective Labels



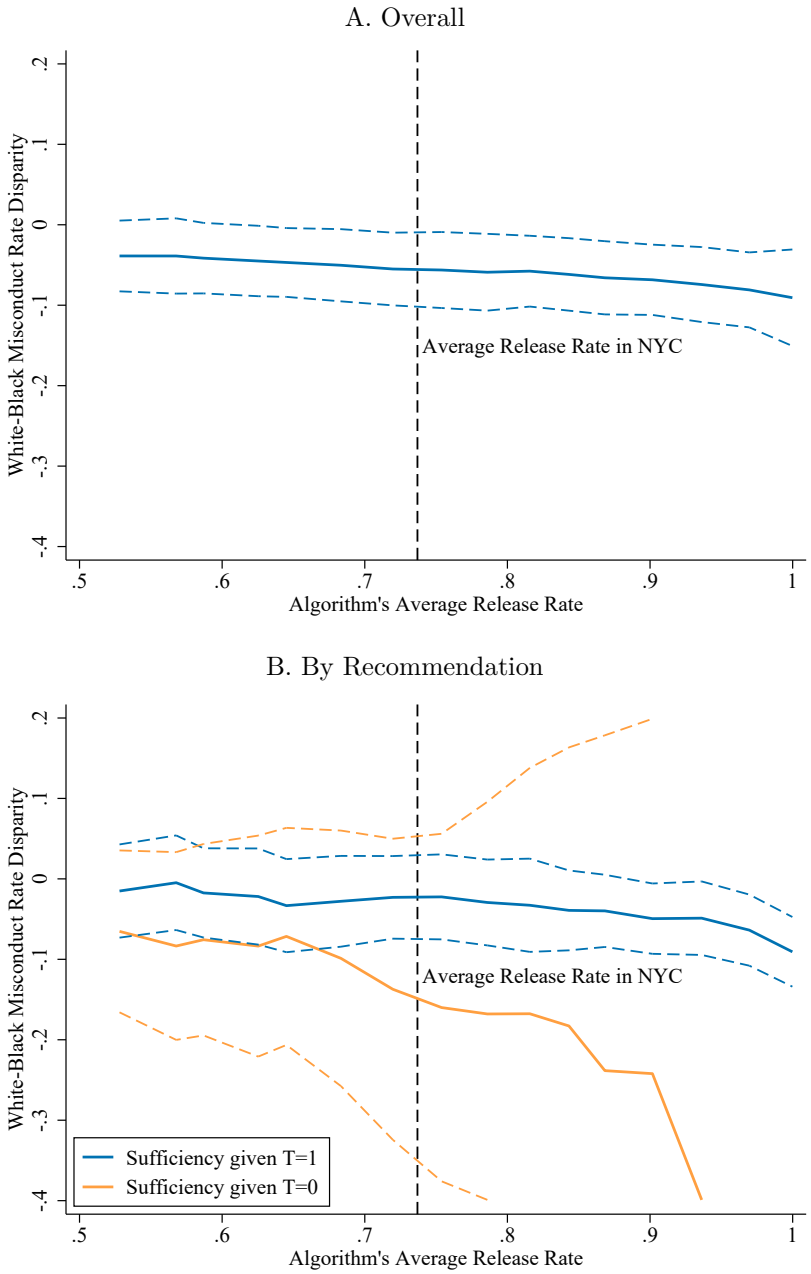
Notes. This figure plots our main discrimination estimates against a potentially biased measure of discrimination that is estimated on the subsample of defendants released before trial. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities from discrimination are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described in the text.

Appendix Figure A6: Decomposition of Algorithmic Discrmination



Notes. This figure plots the range of racial disparities in true and false negative rates, for different average release rates, which make up the disparities due to racial discrimination. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described in the text.

Appendix Figure A7: Measuring Discrimination by Sufficiency of Algorithmic Recommendations



Notes. This figure plots the range of racial disparities in average positive and negative predictive values, or the insufficiency of algorithmic release rate recommendations, for different average release rates. Panel A shows the overall racial disparity while Panel B shows disparities separately by the algorithm's recommendation. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described in the text.

Appendix Table A1: Descriptive Statistics by Sample

	All Defendants		White Defendants		Black Defendants	
	Full Sample	Estimation Sample	Full Sample	Estimation Sample	Full Sample	Estimation Sample
Panel A: Pretrial Release	(1)	(2)	(3)	(4)	(5)	(6)
Released Before Trial	0.852	0.730	0.872	0.767	0.832	0.695
Share ROR	0.601	0.852	0.616	0.852	0.586	0.851
Share Disposed	0.301	0.000	0.274	0.000	0.327	0.000
Share Adjourned	0.191	0.000	0.199	0.000	0.183	0.000
Share Money Bail	0.068	0.144	0.070	0.144	0.066	0.145
Share Other Bail Type	0.332	0.004	0.314	0.004	0.348	0.004
Share Remanded	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: Defendant Characteristics						
White	0.483	0.478	1.000	1.000	0.000	0.000
Male	0.822	0.821	0.831	0.839	0.813	0.804
Age at Arrest	31.819	31.969	31.540	32.055	32.080	31.890
Prior Rearrest	0.192	0.229	0.168	0.204	0.214	0.253
Prior FTA	0.085	0.103	0.071	0.087	0.099	0.117
Panel C: Charge Characteristics						
Number of Charges	1.094	1.150	1.111	1.184	1.078	1.118
Felony Charge	0.184	0.362	0.181	0.355	0.188	0.368
Misdemeanor Charge	0.816	0.638	0.819	0.645	0.812	0.632
Any Drug Charge	0.347	0.256	0.342	0.257	0.352	0.256
Any DUI Charge	0.031	0.046	0.046	0.067	0.017	0.027
Any Violent Charge	0.072	0.143	0.062	0.124	0.081	0.160
Any Property Charge	0.217	0.136	0.209	0.127	0.226	0.144
Cases	1,358,278	595,186	656,711	284,598	701,567	310,588

Notes. This table summarizes the difference between the NYC analysis sample and the full sample of NYC arraignments. The full sample consists of all bail hearings between November 1, 2008 and November 1, 2013. The analysis sample consists of bail hearings that were quasi-randomly assigned to judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records as described in the text. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. ROR (released on Recognizance) is defined as being released without any conditions. FTA (failure to appear) is defined as failing to appear at a mandated court date.

Appendix Table A2: Judge Leniency and Sample Attrition

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Dropped from Sample	0.00007	0.00003	0.00012
	(0.00012)	(0.00013)	(0.00014)
Court x Time FE	Yes	Yes	Yes
Mean Sample Attrition	0.416	0.409	0.424
Cases	1,425,652	726,284	697,597

Notes. This table reports OLS estimates of regressions of judge leniency on an indicator for leaving the sample due to case adjournment or case disposal and court-by-time fixed effects. The regressions are estimated on the sample of all arraignments made in NYC between November 1, 2008 and November 1, 2013. Judge leniency is estimated using data from other cases assigned to a given bail judge. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A3: Tests of Quasi-Random Judge Assignment

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
White	0.00013		
	(0.00009)		
Male	0.00003	0.00003	0.00004
	(0.00014)	(0.00019)	(0.00018)
Age at Arrest	-0.00011	-0.00015	-0.00008
	(0.00004)	(0.00006)	(0.00005)
Prior Rearrest	-0.00021	0.00006	-0.00042
	(0.00011)	(0.00018)	(0.00015)
Prior FTA	0.00016	-0.00011	0.00036
	(0.00016)	(0.00024)	(0.00023)
Number of Charges	-0.00001	-0.00001	-0.00001
	(0.00001)	(0.00001)	(0.00003)
Felony Charge	0.00025	0.00011	0.00039
	(0.00020)	(0.00023)	(0.00025)
Any Drug Charge	-0.00022	-0.00017	-0.00027
	(0.00016)	(0.00021)	(0.00018)
Any DUI Charge	0.00045	0.00051	0.00008
	(0.00027)	(0.00032)	(0.00045)
Any Violent Charge	-0.00008	-0.00023	0.00001
	(0.00023)	(0.00033)	(0.00025)
Any Property Charge	-0.00033	-0.00028	-0.00036
	(0.00018)	(0.00019)	(0.00027)
Joint p-value	[0.10689]	[0.29792]	[0.10136]
Court x Time FE	Yes	Yes	Yes
Cases	595,186	284,598	310,588

Notes. This table reports OLS estimates of regressions of judge leniency on defendant characteristics. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a given bail judge. All regressions control for court-by-time fixed effects. The p-values reported at the bottom of each column are from F-tests of the joint significance of the variables listed in the rows. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A4: First Stage Effects of Judge Leniency

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Judge Leniency	0.960	0.788	1.104
	(0.025)	(0.029)	(0.033)
Court x Time FE	Yes	Yes	Yes
Mean Release Rate	0.730	0.767	0.695
Cases	595,186	284,598	310,588

Notes. This table reports OLS estimates of regressions of an indicator for pretrial release on judge leniency. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a bail judge. All regressions control for court-by-time fixed effects. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

B Econometric Appendix

B.1 Algorithm Estimation Details

This appendix details our baseline algorithmic predictions of pretrial misconduct risk. We use a gradient boosted decision tree model, based on the model that Kleinberg et al. (2018) develop for the NYC pretrial system. We use the same feature set X_i, which includes a total of 38 variables summarizing prior criminal history, charge characteristics, and demographic variables such as the age of the defendant. The outcome variable Y_i is an indicator for pretrial misconduct, defined as either a failure to appear or being arrested for a new crime.

Gradient boosting is an ensemble method that aggregates a number of "weak learners" in an iterative fashion. Here, the weak learners are decision trees, which divide the data through a sequence of binary splits based on the feature set. The algorithm averages multiple decision trees built sequentially on the data, with subsequent iterations up-weighting the observations predicted most poorly by the preceding sequence of trees. The complexity of the gradient boosting algorithm depends on the "depth" of each tree and a "shrinkage" parameter which governs how trees are averaged together.

Following Kleinberg et al. (2018), we choose the model hyperparameters by k-fold cross-validation with five folds. We first select a random 80 percent sample of released defendants, which we take as the training dataset. Applying the cross-validation procedure to this dataset yields an optimal tree depth of 4 and shrinkage parameter of 0.05. We then use the full training set and the remaining 20 percent of released defendants (the test dataset) to fit the gradient boosted decision tree model with these hyperparameters. Finally, we apply the model to the full sample (including defendants detained before trial) to compute risk predictions $p(X_i)$. We use the complete sample to estimate algorithmic discrimination for consistency with Arnold, Dobbie and Hull (2020) and to maximize precision.

B.2 Alternative Regression-Based Risk Predictions

This appendix measures algorithmic discrimination in a simpler regression-based prediction of pretrial misconduct risk, inspired by the Laura and John Arnold Foundation Public Safety Assessment tool (LJAF PSA). The LJAF PSA is used in a number of states and cities to assist bail judges in making pretrial release decisions. LJAF PSA scores are based on nine defendant and case observables: the defendant's age; an indicator for a violent crime charge; an indicator for a pending charge at the time of offense; indicators for a prior misdemeanor, felony, or violent crime conviction; the number of previous failures to appear over the last two years; an indicator for a failure to appear more than two years ago; and an indicator for prior incarceration.

We construct ordinary least squares risk predictions $\hat{Y}_i$ by regressing, in the sample of released defendants, an indicator for pretrial misconduct (either a failure to appear or being rearrested for a new crime) on a set of observed characteristics based on the LJAF PSA inputs. For most characteristics, we are able to match the inputs exactly. We do not observe whether a defendant has a pending charge, however, so we exclude this input. We also do not observe prior incarceration, so we instead use an indicator for prior arrest. As with the main algorithmic predictions, we use these $\hat{Y}_i$ and a range of risk thresholds τ to form release recommendations $T_i = \mathbf{1}[\hat{Y}_i < \tau]$ for all defendants in the sample.

Appendix Figure A3 shows our extrapolation-based estimation of the key race-specific second moments ρ_w and ρ_b when using the regression-based prediction of pretrial misconduct. Appendix

Figure A4 plots the corresponding range of estimated measures of algorithmic discrimination for the regression-based prediction of pretrial misconduct. As with our baseline gradient boosted decision tree algorithm, the regression-based algorithmic recommendations yield similar second-moment estimates for white and Black defendants (of around 0.2) at the average release rate in NYC (73 percent). These estimates and the common mean risk estimates yield a 6.7 percentage point disparity in the recommended release rates of white and Black defendants with the same potential for pretrial misconduct. This discriminatory disparity is a large share (73.6 percent) of the unadjusted release rate disparity in algorithmic recommendations (9.1 percentage points), and a similar share as with our baseline gradient boosted decision tree algorithm. We again find algorithmic discrimination over a wide range of potential release rates, with the estimated Δ statistically distinguishable from zero at all but the highest release rates.

B.3 Alternative Discrimination Measures

This appendix shows how our estimates of race-specific parameters $\{\mu_w, \mu_b, \rho_w, \rho_b\}$ can be used to construct alternative measures of algorithmic discrimination in the NYC pretrial setting. We first estimate race-specific covariances of misconduct potential Y_i^* and algorithmic release recommendations T_i. We then estimate racial disparities in true- and false-negative rates, δ_r^T and δ_r^F, which enter our average discrimination measure Δ. Racial equality in false-negative rates can be seen as satisfying what is known in the computer science literature as "equality of opportunity" (Hardt, Price and Srebro, 2016), meaning that "qualified" white and Black defendants (without pretrial misconduct potential) are released at the same rate. We also show that our estimates can be used to detect departures from what is known in the computer science literature as "sufficiency" (Zafar et al., 2017), and what Kleinberg, Mullainathan and Raghavan (2017) refer to as "calibration," meaning the racial equality of positive and negative predictive values.

Appendix Figure A2 first plots our estimates of race-specific covariances of misconduct potential and algorithmic release recommendations across a range of release rates. These estimates are obtained by $Cov(Y_i^*, T_i \mid R_i) = \rho_{R_i} - \mu_{R_i} \times E[T_i \mid R_i]$. We tend to find a stronger (more negative) covariance for Black defendants than white defendants. This results from the fact that we estimate a higher mean risk μ_{R_i} for Black defendants than for white defendants, while we tend to obtain similar estimates of the second moment ρ_{R_i} and somewhat higher release rates $E[T_i \mid R_i]$ for white defendants.

Appendix Figure A6 next plots our estimates of racial disparities in true- and false-negative rates. These estimates are obtained by the formulas for δ_r^T and δ_r^F in the main text. We find a disparity in false-negative rates that is large and roughly constant across different release rates, where white defendants with misconduct potential tend to be released at a higher rate than Black defendants with misconduct potential. In contrast, the racial disparity in true-negative rates (i.e., the release rate differential among defendants without misconduct potential) is only statistically significantly different from zero at low release rates. These results suggest that a measure of "inequality of opportunity" (that $\delta_w^T - \delta_b^T \neq 0$) could fail to detect overall racial discrimination (that $\Delta \neq 0$) in this setting.

Finally, Appendix Figure A7 plots estimates of algorithmic "insufficiency." Sufficiency is defined by the equality of negative and positive predictive values across race, or equivalently the independence of Y_i^* and R_i given T_i. Paralleling our main discrimination measure Δ, we define insufficiency as:

$$\Sigma = E[E[Y_i^* \mid R_i = w, T_i] - E[Y_i^* \mid R_i = b, T_i]]$$

Here, the inner difference compares the misconduct rate for white and Black defendants, holding fixed the algorithmic recommendation T_i. The outer expectation averages this comparison over the recommendation distribution. A finding of $\Sigma < 0$ indicates white individuals are less risky than Black defendants with identical algorithmic recommendations. As with Δ, this measure can be decomposed:

$$\Sigma = (\sigma_w^R - \sigma_b^R)E[T_i] + (\sigma_w^D - \sigma_b^D)(1 - E[T_i])$$

where $\sigma_r^R = E[Y_i^* \mid R_i = r, T_i = 1]$ is the misconduct rate among released individuals of race r and $\sigma_r^D = E[Y_i^* \mid R_i = r, T_i = 0]$ is the misconduct rate among detained individuals of race r.

To estimate Σ, we use this decomposition and the fact that:

$$\sigma_r^R = 1 - \frac{E[Y_i^* T_i \mid R_i = r]}{E[T_i \mid R_i = r]} = 1 - \frac{\rho_r}{E[T_i \mid R_i = r]}$$
$$\sigma_r^D = 1 - \frac{E[Y_i^*(1 - T_i) \mid R_i = r]}{E[(1 - T_i) \mid R_i = r]} = 1 - \frac{\mu_r - \rho_r}{1 - E[T_i \mid R_i = r]}$$

Panel A of Appendix Figure A7 shows a generally negative Σ across a wide range of algorithmic release rates when we use our estimates of first- and second-moments as inputs to these formulas, meaning that white defendants tend to have lower pretrial misconduct rates than Black defendants conditional on the algorithm's release recommendation. Panel B of Appendix Figure A7 shows that the average misconduct rate disparities in Panel A are driven by large but noisy racial disparities among detained defendants and smaller but more precise racial disparities among released defendants.