Building Non-Discriminatory Algorithms in Selected Data*

David Arnold[†] Will Dobbie[‡] Peter Hull[§]

April 2024

Abstract

We develop new quasi-experimental tools to understand algorithmic discrimination and build non-discriminatory algorithms when the outcome of interest is only selectively observed. These tools are applied in the context of pretrial bail decisions, where conventional algorithmic predictions are generated using only the misconduct outcomes of released defendants. We first show that algorithmic discrimination arises in such settings when the available algorithmic inputs are systematically different for white and Black defendants with the same objective misconduct potential. We then show how algorithmic discrimination can be eliminated by measuring and purging these conditional input disparities. Leveraging the quasi-random assignment of bail judges in New York City, we find that our new algorithms not only eliminate algorithmic discrimination but also generate more accurate predictions by correcting for the selective observability of misconduct outcomes.

[*] We thank Patrick Kline, Ashesh Rambachan, Jonathan Roth, Kaspar Wuthrich, and numerous seminar participants for helpful comments and discussions. We thank Stephanie Lukins and Michelle Wu for outstanding research assistance. We gratefully acknowledge funding through the National Science Foundation (Award #2119849). 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 the DCJS or the OCA. New York State, the DCJS, and the OCA do not assume liability for its contents or use thereof. *JEL*: C26, J15, K42.

[†] University of California, San Diego and NBER. Email: daarnold@ucsd.edu

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

[§] Brown University and NBER. Email: peter_hull@brown.edu

1 Introduction

High-stakes predictive algorithms often generate different predictions for protected groups, even when group identity is omitted from the algorithmic inputs.[1] There are growing concerns that such predictive disparities reflect algorithmic discrimination, defined broadly as the incorporation and perpetuation of systemic biases through the inputs. However, interpreting predictive disparities as algorithmic discrimination is challenging as there may be legitimate differences in the outcome of interest across groups. Such outcomes are also only selectively observed in most high-stakes settings, hampering efforts to both measure and eliminate algorithmic discrimination.

This paper develops quasi-experimental tools to understand algorithmic discrimination and build non-discriminatory algorithms when the outcome of interest is only selectively observed. We apply these tools in the context of pretrial bail decisions, where conventional algorithmic predictions are generated using only the misconduct outcomes of released defendants and may also rely on biased inputs. Judges are mandated by law to allow most defendants to be released before trial while minimizing the risk of pretrial misconduct, and therefore risk violating U.S. anti-discrimination law if they release white and Black defendants with the same objective misconduct potential at different rates. We thus define algorithmic discrimination as racial disparities in risk score predictions conditional on misconduct potential. This definition is consistent with the legal objective of bail judges and recent analyses of discrimination in both computer science (e.g., Zafar et al., 2017; Berk et al., 2021) and economics (e.g., Arnold, Dobbie and Hull, 2022; Bohren, Hull and Imas, 2023; Baron et al., Forthcoming).

We focus our analysis on linear prediction models, which include the most common algorithms in the pretrial setting and elsewhere.[2] Our first contribution is to show how algorithmic discrimination arises in such models from disparities in the algorithmic inputs among individuals with the same outcome of interest. For example, consider a scenario where Black defendants have more prior criminal convictions than white defendants with the same objective misconduct potential due to racial bias in past policing and prosecutorial decisions. Algorithmic predictions that put positive weight on this input would then, all else equal, generate systematically higher risk scores for Black defendants than white defendants with the same misconduct potential. We propose a new graphical tool to assess the net effect of such conditional disparities on algorithmic discrimination.

This framework yields a conceptually simple way to build non-discriminatory algorithms: use an initial pre-processing step to purge conditional disparities in the inputs, thereby eliminating algorithmic discrimination at its source. We first show how this approach can be easily implemented when the outcome of interest is fully observable. Consider a scenario where we observe the true pretrial misconduct potential for all past defendants. We can regress each algorithmic input on race while controlling for pretrial misconduct potential in this sample. By subtracting the race component of these regressions from each algorithmic input, we create pre-processed inputs that have no conditional input disparities by construction. We can then use these pre-processed inputs to build a non-discriminatory algorithm using, for example, a linear regression of true misconduct potential on the pre-processed

[1] For example, there are large protected-group disparities in the predictions generated by algorithms used to make hiring decisions (Li, Raymond and Bergman, 2020), educational tracking decisions (Bergman, Kopko and Rodriguez, 2023), child protective services decisions (Chouldechova et al., 2018; Cheng et al., 2022), health care decisions (Obermeyer et al., 2019; Nguyen et al., 2021; Gillette et al., 2022), pretrial release decisions (Skeem and Lowenkamp, 2016; Arnold, Dobbie and Hull, 2021), lending decisions (Fuster et al., 2022), and tax auditing decisions (Elzayn et al., 2023).

[2] In the pretrial setting, both the Arnold Ventures Public Safety Assessment (PSA) tool and the New York City pretrial system use linear models to generate algorithmic predictions. Many medical risk scores also use linear models, such as the HEART score for predicting cardiac events and the Apgar score for evaluating newborn health.

inputs.[3] We also discuss alternative in-processing and post-processing solutions: in-processing constrains the model to balance the conditional input disparities, while post-processing directly removes algorithmic discrimination from the conventional model predictions.

The key empirical challenge in applying this framework is that the outcome of interest is only selectively observed in most high-stakes settings. For example, in the pretrial context, misconduct outcomes are observed only among past defendants who are released before trial and not among defendants who are detained. Our second contribution is to show how this challenge can be overcome with experimental or quasi-experimental variation. We first show that the challenge reduces to estimating a small set of moments capturing the mean of the selectively observed outcome and its correlation with race and the non-race algorithmic inputs. We then show how these key moments can be estimated by "selection-correcting" the observed misconduct potential mean and correlations using exogenous shocks to the selection mechanism.

We apply these new tools in the New York City pretrial system, using the quasi-random assignment of bail judges to address the selective observability of pretrial misconduct outcomes. We consider algorithmic inputs that are common in a range of pretrial risk assessments, many of which include information about prior criminal history and characteristics of the current charge. We find that all of these inputs systematically differ across white and Black defendants with the same objective misconduct potential, with nearly all of these inputs leading to higher risk scores for Black defendants. For example, we find Black defendants are significantly more likely to have a prior felony conviction than white defendants conditional on misconduct potential. Prior felony convictions are also positively correlated with misconduct potential, so algorithms that put positive weight on this input generate higher risk scores for Black defendants than white defendants with the same misconduct potential.

Overall, we find that a conventional algorithm generates predicted risk scores that are 2.5 percentage points (8%) higher for Black defendants than white defendants conditional on true misconduct potential. Our non-discriminatory algorithm purges the conditional input disparities, thereby eliminating the risk score disparity between white and Black defendants with the same misconduct potential. By correcting for the selective observability of misconduct outcomes, our non-discriminatory algorithm also generates more accurate risk score predictions—reducing mean squared error by 12.1% compared to the conventional model. Our pre-processing adjustment thus offers a rare "free lunch" in this setting, improving both fairness and accuracy compared to conventional models.[4]

This paper builds on a recent empirical literature studying the performance and implementation of algorithms in high-stakes settings (e.g., Cowgill, 2018; Bhatt et al., 2024). Most relatedly, Kleinberg et al. (2018) use quasi-random judge assignment to show that predictive algorithms in the pretrial context can both reduce racial disparities and improve accuracy relative to existing judges. Other recent work studies the implementation of algorithms alongside human decision-makers in other contexts, finding mixed impacts on both racial disparities and accuracy (e.g., Stevenson, 2018; Stevenson and Doleac, Forthcoming; Rittenhouse, Putnam-Hornstein and Vaithianathan, 2023; Grimon and Mills, 2022; Albright, 2024). Our analysis shows how quasi-experimental variation can be used to construct

[3] Conventional pre-processing adjustments residualize each algorithmic input on race *without* controlling for true misconduct potential. These adjustments generally fail to purge conditional input disparities unless race is uncorrelated with true misconduct potential, which is unlikely in most high-stakes settings (e.g., Arnold, Dobbie and Hull, 2022).

[4] We consider several extensions, finding similar results when generating binary release recommendations at the judges' current release rates, considering different misconduct types or extrapolation approaches, or building non-discriminatory algorithms using only released defendants. For example, we find that the risk score disparity translates into a 5% lower suggested release rate for Black defendants versus white defendants with the same objective misconduct potential.

non-discriminatory algorithms while accounting for human decisions and other important institutional contexts that generate selectively observed outcomes.

Our analysis also builds on a large theoretical literature studying sources and definitions of algorithmic discrimination. This literature shows how biased data can lead to different types of algorithmic disparities (e.g., Kallus and Zhou, 2018; Rambachan and Roth, 2020; Madras et al., 2019), how different notions of discrimination are related (e.g., Chouldechova, 2017; Kleinberg, Mullainathan and Raghavan, 2017; Liang, Lu and Mu, 2023), and how different types of disparities can be purged using pre-processing adjustments (e.g., Kamiran and Calders, 2012; Feldman et al., 2015; Calmon et al., 2017; Pope and Sydnor, 2011), in-processing adjustments (e.g., Zafar et al., 2017; Woodworth et al., 2017; Mishler and Kennedy, 2022), and post-processing adjustments (e.g., Hardt, Price and Srebro, 2016; Pleiss et al., 2017; Kim, Ghorbani and Zou, 2019; Mishler, Kennedy and Chouldechova, 2021). We contribute to this literature by developing a tractable regression-based framework to understand and eliminate algorithmic discrimination. Our framework yields intuitive graphical representations and simple regression calculations for understanding and eliminating algorithmic discrimination in an ideal scenario in which the outcome of interest is fully observable.

Finally, this paper adds to a recent literature showing how algorithmic discrimination and accuracy can be studied in settings when the outcome of interest is only selectively observed—sometimes known as the "selective labels" problem (Lakkaraju et al., 2017; Kleinberg et al., 2018). Much of this literature uses selection-on-observables assumptions (e.g., Schulam and Saria, 2017; Coston et al., 2020), with some work bounding deviations from selection-on-observables via sensitivity models (e.g., Rambachan, Coston and Kennedy, 2023). We add to this literature by showing how, in our regression-based framework, the selection challenge reduces to the challenge of estimating a small number of moments with quasi-experimental variation. Our estimation approach can be seen as a version of "identification at infinity" in conventional sample selection models (e.g., Chamberlain, 1986; Heckman, 1990), using indicators for quasi-randomly assigned judges as instrumental variables (IVs). This approach builds on recent advances in measuring the fairness and accuracy of human decision-makers using quasi-random examiner assignments (Arnold, Dobbie and Hull, 2021, 2022; Angelova, Dobbie and Yang, 2023; Chan, Gentzkow and Yu, 2022; Baron et al., Forthcoming).[5] Such assignment variation is already widely used, suggesting our new tools may be applied in many other settings like lending decisions, medical diagnoses, and foster care placement.

2 Framework

We develop a framework to study algorithmic discrimination when the outcome of interest is only selectively observed, focusing on linear models. We use this framework to show how algorithmic discrimination arises from disparities in the algorithmic inputs among individuals with the same outcome of interest. We then show how experimental or quasi-experimental variation can be used to measure and purge such disparities, letting us build non-discriminatory algorithms in selected data.

[5] Importantly, this approach does not require a model of decision-making or the usual IV monotonicity assumption, which may be violated in examiner IV designs (e.g., Frandsen, Lefgren and Leslie, 2023). We only require that the available quasi-experiment induces enough variation in the selection mechanism to reliably estimate the key moments.

2.1 Setting

Consider a population of individuals indexed by i and distinguished by three variables: $G_i \in \{0,1\}$, $X_i \in \mathbb{R}^K$, and $Y_i^* \in \mathbb{R}$. Here G_i denotes membership in a protected group, X_i is a vector of K observable characteristics, and Y_i^* is an outcome of interest. X_i includes a constant and may or may not include G_i. Both G_i and X_i are observed by the econometrician, but Y_i^* is only selectively observed among some set of individuals indicated by $D_i = 1$.

To make this setting concrete, consider a population of white and Black defendants at a pretrial hearing, where $G_i = 1$ indicates that defendant i is Black and X_i includes additional demographic information, information from the defendant's criminal records, and various charge characteristics. Bail judges are legally mandated to minimize the risk of pretrial misconduct, such as a failure to appear in court (FTA) or a rearrest for a new crime, while releasing most defendants. A defendant's pretrial misconduct potential Y_i^* is thus the outcome of interest. Bail judges may use the observable characteristics in X_i to determine which defendants to release before trial, as well as private information from a short pretrial hearing. If a defendant is released ($D_i = 1$), then pretrial misconduct potential is realized. Otherwise, a defendant is detained ($D_i = 0$), and Y_i^* is unobserved.

An algorithm uses the observable characteristics in X_i (i.e., the algorithmic inputs) to predict Y_i^*. We focus our analysis on linear predictions $\hat{Y}_i = X_i'\beta$, where the algorithm is parameterized by a coefficient vector $\beta \in \mathbb{R}^K$. The set of linear models includes the most widely-used pretrial risk score algorithms, as well as algorithms in a variety of other high-stakes settings. Linearity is common in practice because it makes predictive algorithms easier to implement and interpret.

We measure algorithmic discrimination using protected-group disparities in the algorithmic predictions $\hat{Y}_i$ among defendants with the same objective misconduct potential Y_i^*. This measure follows recent work in both computer science (e.g., Zafar et al., 2017; Berk et al., 2021) and economics (e.g., Arnold, Dobbie and Hull, 2022; Bohren, Hull and Imas, 2023; Baron et al., Forthcoming) and is linked to the legal theory of disparate impact (Arnold, Dobbie and Hull, 2022). Our measure is also consistent with the judges' legal objective, as judges risk violating U.S. anti-discrimination law if they release white and Black defendants with the same objective misconduct potential at different rates.[6]

Formally, we measure algorithmic discrimination by the coefficient Δ in the linear regression:

$$\hat{Y}_i = \alpha + \Delta G_i + \phi Y_i^* + \varepsilon_i. \tag{1}$$

Here, Δ captures disparities in the algorithmic predictions $\hat{Y}_i$ among defendants with the same objective misconduct potential Y_i^*. Since higher pretrial risk scores generally lead to harsher treatments, $\Delta > 0$ can be interpreted as discrimination against the protected group (e.g., Black defendants) while $\Delta < 0$ can be interpreted as discrimination against the non-protected group (e.g., white defendants).

The main identification challenge is that Y_i^* is only selectively observed among defendants with $D_i = 1$, so Δ cannot be directly estimated by running this regression in the full population. Estimating Equation (1) in the $D_i = 1$ subpopulation is likely to yield a biased estimate of Δ since D_i is likely correlated with (G_i, Y_i^*). We return to this challenge and our solution in Section 2.4 after considering how algorithmic discrimination can be understood and eliminated if Y_i^* were fully observable.

[6] Our approach to constructing non-discriminatory algorithms extends to other discrimination measures based on Y_i^*. Our in-processing solution discussed below can be applied to any non-discrimination constraint based on moments of $\{G_i, X_i, Y_i^*\}$, and versions of our quasi-experimental strategy can be used to estimate these moments.

2.2 Understanding Algorithmic Discrimination

Algorithmic discrimination Δ arises from disparities in the algorithmic inputs among defendants with the same objective misconduct potential. Formally, by Equation (1) and the linearity of $\hat{Y}_i = X_i'\beta$:

$$\Delta = \sum\nolimits_k \Gamma_k \beta_k, \tag{2}$$

where Γ_k are coefficients from regressions of the form:

$$X_{ik} = \mu_k + \Gamma_k G_i + \psi_k Y_i^* + \nu_{ik}. \tag{3}$$

As in Equation (1), the Γ_k coefficients capture disparities in each algorithmic input among defendants with the same misconduct potential. Equation (2) shows that algorithmic discrimination is a linear combination of these conditional input disparities, weighted by the algorithm's coefficients β_k.

To make Equation (2) concrete, consider an example in which Black defendants have more prior criminal convictions (X_{ik}) than white defendants with the same objective misconduct potential (Y_i^*) due to racial bias in past policing and prosecutorial decisions. Here, $\Gamma_k > 0$. Equation (2) shows that algorithmic predictions that put positive weight on the number of prior criminal convictions ($\beta_k > 0$) would then, all else equal, generate systematically higher risk scores for Black defendants than white defendants with the same misconduct potential. That is, $\Gamma_k \beta_k > 0$. Intuitively, if Black defendants have more prior criminal convictions conditional on latent misconduct potential, then they must receive higher risk scores conditional on latent misconduct potential. The magnitude of such algorithmic discrimination depends on both the conditional disparity in the number of prior criminal convictions (Γ_k) and the weight associated with that factor in the model (β_k).

Panel A of Figure 1 builds on this example, illustrating a new graphical tool for understanding the drivers of algorithmic discrimination. Each point in the figure represents a hypothetical algorithmic input k with its coefficient β_k plotted against its conditional disparity Γ_k. When $\Gamma_k \beta_k > 0$ (first and third quadrants), input k contributes positively to algorithmic discrimination. When $\Gamma_k \beta_k < 0$ (second and fourth quadrants), input k contributes negatively to algorithmic discrimination. In both cases, the magnitude of input k's contribution to algorithmic discrimination is given by the area connecting input k's point to the origin, $|\Gamma_k \beta_k|$. In Panel A of Figure 1, the sum of each input's contribution in the first and third quadrants is greater in magnitude than that in the second and fourth quadrants, resulting in algorithmic discrimination, on net, against the protected group.

In principle, points could lie in different quadrants such that $\sum_k \Gamma_k \beta_k = 0$ despite $\Gamma_k \neq 0$. Panel B of Figure 1 shows such an example, where the conditional disparities in algorithmic inputs Γ_k are small and uncorrelated with the algorithmic coefficients β_k. In this case, the conditional input disparities Γ_k cancel out, resulting in no algorithmic discrimination ($\Delta = 0$). In practice, such balancing-out is unlikely to arise by chance in conventional algorithmic constructions. However, Panel B of Figure 1 suggests we can eliminate algorithmic discrimination using alternative constructions. If all inputs were adjusted to have an exactly zero conditional disparity, i.e., $\Gamma_k = 0$, then any linear algorithm would exhibit zero algorithm discrimination.

2.3 Building a Non-Discriminatory Algorithm

Our non-discriminatory algorithm uses an initial pre-processing adjustment that partially residualizes each algorithmic input using the race component of Equation (3), such that each input has an exactly zero conditional disparity. Formally, define:

$$\tilde{X}_{ik} = X_{ik} - \Gamma_k G_i. \tag{4}$$

By construction, the pre-processed $\tilde{X}_{ik}$ exhibit no conditional disparity among individuals with the same Y_i^*. That is, the coefficient from regressing each $\tilde{X}_{ik}$ on G_i, controlling for Y_i^*, is mechanically zero. It follows from Equation (2) that any algorithmic prediction that is linear in these $\tilde{X}_{ik}$ exhibits no algorithmic discrimination. Intuitively, pre-processing eliminates algorithmic discrimination by removing any variation along the horizontal axis in plots like Panel A of Figure 1.[7]

We can then build a non-discriminatory algorithm by regressing true misconduct potential on the pre-processed inputs. Formally, let $\tilde{\beta} = E[\tilde{X}_i \tilde{X}_i']^{-1} E[\tilde{X}_i Y_i^*]$ where $\tilde{X}_i$ collects the pre-processed $\tilde{X}_{ik}$, and let $\hat{Y}_i^{Pre} = \tilde{X}_i' \tilde{\beta}$ denote the resulting pre-processed algorithmic predictions. Here, $\tilde{\beta}$ minimizes mean squared error (MSE): $\tilde{\beta} = \arg\min_b E\left[(Y_i^* - \tilde{X}_i' b)^2\right]$.

Two alternative non-discriminatory algorithm constructions build on other adjustments from the literature to balance-out algorithmic discrimination such that $\sum_k \Gamma_k \beta_k = 0$ despite $\Gamma_k \neq 0$. First consider an in-processing procedure that directly minimizes the MSE of algorithmic predictions while constraining $\Delta = 0$. Formally, let $\hat{Y}_i^{In} = X_i' \beta^*$ denote the in-processed algorithmic predictions where

$$\beta^* = \arg\min_b E\left[(Y_i^* - X_i' b)^2\right] \text{ s.t. } \Gamma' b = 0, \tag{5}$$

for $\Gamma = (\Gamma_1, \ldots, \Gamma_K)'$. Equation (5) is a constrained least-squares problem that minimizes the MSE of linear predictions based on the original X_i, subject to the constraint of no algorithmic discrimination. The solution to this problem is:

$$\beta^* = \beta^U - E[X_i X_i']^{-1} \Gamma \left(\Gamma' E[X_i X_i']^{-1} \Gamma\right)^{-1} \Gamma' \beta^U, \tag{6}$$

where $\beta^U = E[X_i X_i']^{-1} E[X_i Y_i^*]$ is the unconstrained solution to Equation (5). Intuitively, β^* projects the unconstrained coefficient vector β^U onto the vector of conditional input disparities Γ (weighting by $E[X_i X_i']^{-1}$) and takes the residuals. By construction, these residuals balance-out algorithmic discrimination by removing any trend through the origin in plots like Panel A of Figure 1. For example, $\Delta > 0$ if the slope of the line of best fit through the origin is positive.

The second alternative construction is a post-processing procedure that subtracts the average level of algorithmic discrimination from the predictions of the protected group. Formally, consider $\hat{Y}_i^{Post} = X_i' \beta^U - \Delta^U G_i$, where Δ^U is the level of algorithmic discrimination in the unconstrained algorithm (i.e., $\Gamma' \beta^U$). By construction, the coefficient from regressing $\hat{Y}_i^{Post}$ on G_i controlling for Y_i^* is zero. Intuitively, this is because post-processing adds to the algorithmic inputs a new component $X_{i,K+1} = G_i$ where $\Gamma_{K+1} = 1$ and $\beta_{K+1} = -\Delta^U$. This adds a new point to plots like Panel A of Figure 1, balancing out discrimination because $\sum_{k=1}^{K+1} \Gamma_k \beta_k = \Delta^U + 1 \cdot (-\Delta^U) = 0$.

[7] By comparison, pre-processing adjustments that do not control for Y_i^* generally fail to eliminate algorithmic discrimination because of omitted variables bias. Formally, consider an algorithm constructed from $X_{ik} - \Lambda_k G_i$ instead of $\tilde{X}_{ik}$, where Λ_k comes from the regression: $X_{ik} = \tau_k + \Lambda_k G_i + \upsilon_{ik}$. From Equation (3), we have that $\Lambda_k = \Gamma_k + \psi_k \left(E[Y_i^* \mid G_i = 1] - E[Y_i^* \mid G_i = 0]\right)$, such that $\Gamma_k \neq \Lambda_k$ when $\psi_k \neq 0$ and $E[Y_i^* \mid G_i = 1] \neq E[Y_i^* \mid G_i = 0]$.

We focus on the pre-processed algorithm as it has at least two practical advantages. First, the computation of $\hat{Y}_i^{Pre}$ is relatively simple and transparent with a two-step procedure that involves only linear regressions. Second, pre-processing the algorithmic inputs ensures no discrimination in *any* linear algorithm based on these inputs. However, there are settings in which in-processed and post-processed algorithms may be preferred. In-processing yields the most accurate non-discriminatory predictions based on (X_i, G_i), while post-processing can eliminate discrimination from non-linear algorithmic predictions. We apply and compare all three algorithms below.

Each solution is straightforward to implement when Y_i^* is fully observable. For example, Equations (2) and (4) show that both the level of discrimination in a given algorithm and the pre-processing step that eliminates algorithmic discrimination are simple functions of Γ, where Γ is a known function of the first and second moments of $\{X_i, G_i, Y_i^*\}$. If we can estimate these moments, we can measure and eliminate algorithmic discrimination (the same logic holds for the in-processing and post-processing solutions). The main identification challenge is that Y_i^* is only selectively observed in the subpopulation of individuals with $D_i = 1$, so the key moments cannot be directly estimated.

2.4 Identification

We next show that the selection challenge can be overcome by estimating a small set of moments that captures the mean of the selectively observed outcome and its correlation with race and the non-race algorithmic inputs. To see this, note that each element of Γ can be written:

$$\Gamma_k = \frac{E[(G_i - \kappa - \gamma Y_i^*)X_{ik}]}{E[(G_i - \kappa - \gamma Y_i^*)^2]} = \frac{E[G_i X_{ik}] - \kappa E[X_{ik}] - \gamma E[Y_i^* X_{ik}]}{E[G_i^2] + \kappa^2 + \gamma^2 E[Y_i^{*2}] - 2\left(\kappa E[G_i] + \kappa\gamma E[Y_i^*] + \gamma E[G_i Y_i^*]\right)},$$

where $\gamma = \frac{Cov(Y_i^*, G_i)}{Var(Y_i^*)} = \frac{E[G_i Y_i^*] - E[G_i]E[Y_i^*]}{E[Y_i^{*2}] - E[Y_i^*]^2}$, $\kappa = E[G_i] - \gamma E[Y_i^*]$, and $E[G_i^2] = E[G_i]$ since G_i is binary. It follows that Γ is identified by the $(3K + 4) \times 1$ moment vector:

$$\Theta = [\underbrace{E[G_i], E[X_i]', E[G_i X_i]'}_{\text{Directly estimable}}, \underbrace{E[Y_i^*], E[Y_i^{*2}], E[G_i Y_i^*], E[X_i' Y_i^*]}_{\text{Selection-affected}}]'. \tag{7}$$

The first $2K + 1$ elements of Θ do not involve the selectively observed Y_i^* and can thus be directly estimated. The selection challenge therefore reduces to that of estimating the remaining $K + 3$ elements of Θ involving Y_i^*. When Y_i^* is binary, the challenge further simplifies to estimating only $K + 2$ selection-affected moments since $E[Y_i^{*2}] = E[Y_i^*]$.

We can estimate the selection-affected moments by selection-correcting their observed analogs using experimental or quasi-experimental variation in the selection mechanism. To build intuition for this approach, consider an experiment in which a researcher randomly assigns individuals to an intervention that directly reveals Y_i^*. Formally, let $Z_i \in \{0, 1\}$ indicate randomized assignment to the intervention and let $D_i(z) \in \{0, 1\}$ indicate the potential observability of Y_i^* when $Z_i = z$ such that $D_i = D_i(Z_i)$. Suppose $D_i(1) = 1$, such that Y_i^* is observed for all individuals assigned to the intervention. Then, any moment of the form $E[h(X_i, G_i, Y_i^*)]$, for some $h(\cdot)$, is directly estimable by the observed analog in the subpopulation of individuals with $D_i = Z_i = 1$:

$$E[h(X_i, G_i, Y_i^*) \mid D_i = 1, Z_i = 1] = E[h(X_i, G_i, Y_i^*) \mid D_i(1) = 1] = E[h(X_i, G_i, Y_i^*)]. \tag{8}$$

Here, the first equality follows from the randomization of Z_i while the second equality follows from the fact that $D_i(1) = 1$. Equation (8) thus shows how to estimate the selection-affected moments in Θ when such a randomized intervention is possible.

The selection-affected moments in Θ can also be estimated by extrapolating variation in their observed analogs across an as-good-as-randomly assigned $Z_i \in \mathcal{Z}$. Formally, suppose Z_i is as-good-as-randomly assigned and only affects observed outcomes through the selection mechanism. These assumptions make Z_i independent of $\{X_i, G_i, D_i(\cdot), Y_i^*\}$, where we again write $D_i(z)$ as the potential observability of Y_i^* when $Z_i = z$ such that $D_i = D_i(Z_i)$. Define:

$$\pi_z = Pr(D_i = 1 \mid Z_i = z) = Pr(D_i(z) = 1) \tag{9}$$

and

$$\rho_z^h = E[h(X_i, G_i, Y_i^*) \mid D_i = 1, Z_i = z] = E[h(X_i, G_i, Y_i^*) \mid D_i(z) = 1], \tag{10}$$

where the second equalities in both Equation (9) and Equation (10) follow from the independence of Z_i, as before. Both π_z and ρ_z^h are directly estimable. We can obtain estimates of the unselected $E[h(X_i, G_i, Y_i^*)]$ by extrapolating variation in the estimated ρ_z^h across values of z, in the direction of values where the estimated π_z is close to one.

Empirically, we estimate the required moments in Θ by extrapolating the ρ_j^h estimates across a set of as-good-as-randomly assigned judges indexed by j. This approach to estimating the required moments is conceptually similar to how average potential outcomes at a treatment cutoff can be extrapolated from nearby observations in a regression discontinuity design (particularly in "donut" designs where the data in some window of the treatment cutoff are excluded). In our setting, the required moments are extrapolated from as-good-as-randomly assigned judges with release rates π_j close to the "cutoff" of 1. For example, estimates of the required moments may come from the vertical intercept at release rate of 1 of linear or local linear regressions of the averages of $h(X_i, G_i, Y_i^*)$ among released defendants on estimated release rates across judges. This approach to estimating the algorithmic counterfactual at a given release threshold is closely related to that described in Hull (2020) and Arnold, Dobbie and Hull (2022), who consider different extrapolations of quasi-experimental moments in the spirit of "identification at infinity" in sample selection models.[8]

3 Application

3.1 The NYC Pretrial System

We apply our tools in the NYC pretrial system, which is one of the largest in the country. Bail judges are legally mandated to release most criminal defendants while minimizing the risk of pretrial misconduct. Judges are granted considerable discretion in determining which defendants should be released, but they cannot discriminate against minorities and other protected groups even when group identity contains information about the risk of criminal misconduct (Yang and Dobbie, 2020).

[8] One important advantage of this approach is that it can be justified without a conventional first-stage monotonicity assumption: our extrapolated parameters are valid as long as the *average* relationship between conditional misconduct rates and release rates across judges can be reliably estimated. This is likely to hold given the large number of judges with high release rates.

To help guide their decisions, many jurisdictions give bail judges an algorithmic risk assessment that predicts each defendant's likelihood of misconduct and recommends whether to release or detain the defendant. The NYC tool generates risk scores using a linear model in which each algorithmic input is associated with a given point value. The overall risk score is then mapped to a specific release recommendation, which ranges from recommend for release with no conditions for observably low-risk defendants to not recommended for release for observably high-risk defendants.

We exploit two features of the pretrial system. First, pretrial misconduct potential is observed among the selected subset of defendants that judges choose to release before trial. Misconduct potential is, however, unobserved among defendants detained before trial, so we cannot directly estimate the required moments to measure and eliminate algorithmic discrimination. Second, the case assignment procedures used in the NYC pretrial system generate quasi-random variation in bail judge assignment for defendants arrested at the same time and place. This variation in judge assignment generates quasi-experimental variation in the probability a defendant is released before trial. Appendix Table A1 confirms that judge assignment to cases is balanced on all observable defendant and case characteristics conditional on court-by-time fixed effects, and Appendix Table A2 shows that judge assignment has a strong first-stage effect on the probability that a defendant is released before trial.[9]

Appendix B gives additional institutional details—including the specific release conditions, the information available to NYC judges, and details of bail judge assignment.

3.2 Data and Summary Statistics

We observe the universe of arraignments in NYC between November 1, 2008 and November 1, 2013. The data contain information on the defendant's sex, race, date of birth, and county of arrest, as well as the (anonymized) identity of the assigned bail judge. We categorize defendants as white (including both non-Hispanic and Hispanic white individuals) or Black (non-Hispanic individuals). The data also contain information on each defendant's current offense, whether the defendant was released before trial, their history of prior criminal convictions, and their history of pretrial misconduct (both FTA and rearrests).[10] Appendix B provides additional details on the sample construction.

Panels A and B of Table 1 summarize the sample. Overall, 72% of defendants are released before trial (76% of white defendants and 69% of Black defendants), and 30% of released defendants later engage in pretrial misconduct (27% of white defendants and 34% of Black defendants). A large share of defendants have been previously convicted of a misdemeanor or felony, charged with a drug offense, or previously rearrested while on pretrial release—all with higher shares among Black defendants.

Panel C of Table 1 summarizes pretrial risk scores from a conventional model that regresses an indicator for pretrial misconduct (Y_i^*) on the 12 non-race characteristics (X_{ik}) in Panel B of Table 1 in the sample of defendants who are released before trial ($D_i = 1$). These X_{ik} are common in a range of pretrial risk assessments, many of which include information about prior criminal history and characteristics of the current charge. We then predict risk scores ($\hat{Y}_i$) for all defendants in the sample. The conventional model predicts that 31% of white defendants and 34% of Black defendants engage in pretrial misconduct if released, showing that Black defendants are observably riskier than

[9] Our analysis also relies on an exclusion restriction, namely that judges can only systematically affect pretrial misconduct outcomes through pretrial release decisions. This assumption is standard in the literature (e.g., Arnold, Dobbie and Yang, 2018; Dobbie, Goldin and Yang, 2018; Arnold, Dobbie and Hull, 2022; Ouss and Stevenson, 2023), where it is supported by empirical tests.

[10] We take either form of pretrial misconduct as the primary outcome but explore robustness to other measures.

white defendants. We also see that the average risk score prediction for white and Black defendants is higher than the observed pretrial misconduct rate among released white and Black defendants, respectively. This reflects the fact that detained defendants are observably riskier than released defendants. We next apply our framework to determine the extent to which this disparity reflects algorithmic discrimination, while also addressing this selection challenge.

3.3 Results

Understanding Algorithmic Discrimination. Panel C of Figure 1 plots estimates of the conditional disparities Γ_k and algorithmic coefficients β_k for each of the 12 non-race inputs in the conventional model, illustrating our new graphical tool for understanding the drivers of algorithmic discrimination. We obtain the estimates of Γ_k in three steps. In the first step, we estimate the key selection-affected moments in Equation (7)— π_j and ρ_j^h for each $h(\cdot)$ and each bail judge j—with ordinary least square (OLS) regressions of the form:

$$D_i = \sum_j \pi_j Z_{ij} + C_i'\gamma + \varepsilon_i \tag{11}$$

and

$$h(X_i, G_i, Y_i^*) = \sum_j \rho_j^h Z_{ij} + C_i'\lambda^h + \nu_i^h, \tag{12}$$

where the second regression is estimated on the subpopulation of released defendants. Here, D_i again indicates whether individual i is released before trial, X_i is the vector of non-race characteristics shown in Panel B of Table 1, G_i indicates whether a defendant is Black, and Y_i^* is an indicator for pretrial misconduct among released defendants. The Z_{ij} dummies indicate assignment to each judge j, and C_i is a de-meaned vector of controls (court-by-time fixed effects), such that π_j and ρ_j^h capture regression-adjusted release rates and selected moments for each judge j.[11] In the second step, we estimate the unselected moments $E[h(X_i, G_i, Y_i^*)]$ by extrapolating the $\hat{\rho}_j^h$ estimates towards judges with high $\hat{\pi}_j$ estimates.[12] Finally, we plug our estimates of the unselected moments and sample analogs of the directly estimable moments in Equation (7) into the formula for Γ_k.

Panel C of Figure 1 shows that all of our algorithmic inputs systematically differ for white and Black defendants with the same objective misconduct potential. All estimates of Γ_k are statistically distinguishable from zero except for the one corresponding to the constant term, which trivially has no conditional disparity.[13] The figure also shows that nearly all of these inputs lead to higher risk score predictions for Black defendants compared to white defendants. Most of the estimated (Γ_k, β_k) points are in the first and third quadrants, contributing to algorithmic discrimination against Black defendants in the conventional model. For example, conditional on misconduct potential, we find that Black defendants are significantly more likely to have a prior felony conviction than white

[11] Formally, $\pi_j = Pr(D_i(j) = 1)$ and $\rho_j^h = E[h(X_i, G_i, Y_i^*) \mid D_i(j) = 1]$ when judges are quasi-randomly assigned (such that $Z_i = \sum_j \pi_j Z_{ij}$ is independent of (X_i, G_i, Y_i^*) given C_i) and when regression-adjustment for C_i is sufficient. The latter condition holds, for example, when release rates and released misconduct rates are linear in the court-by-time effects for each judge and defendant race. See Section IV.A of Arnold, Dobbie and Hull (2022) for additional details.

[12] Following Arnold, Dobbie and Hull (2022), our baseline extrapolation uses a flexible local linear regression regression that weights inversely by the estimated variance of $\hat{\rho}_j^h$. Appendix Figure A1 plots example extrapolations for $E[Y_i^*]$, $E[G_i Y_i^*]$, and $E[X_{ik} Y_i^*]$ where X_{ik} indicates any prior felony conviction.

[13] The p-value for testing $\Gamma_k = 0$ is less than 0.01 for all inputs except for Any Drug Charge, where $p = 0.07$.

defendants ($\Gamma_k > 0$). Having a prior felony conviction is also positively correlated with misconduct potential ($\beta_k > 0$). Therefore, this input generates higher risk scores for Black defendants than white defendants with the same objective misconduct potential since $\Gamma_k \beta_k > 0$.

The horizontal axis of Figure 2 summarizes the net effect of the conditional input disparities, finding $\Delta = 0.03$ (SE<0.01) for the conventional model. To capture the accuracy of the algorithm, we plot the mean squared error (MSE), derived from estimates of the moments in Equation (7), on the vertical axis.[14] The MSE of the conventional model is 0.24.

Eliminating Algorithmic Discrimination. Before turning to our non-discriminatory algorithms, we first explore the impact of our selection-correction procedure alone. In Figure 2, we find that the MSE decreases by about 12.5% in the unselected model relative to the conventional model, implying significant gains in accuracy. We find minimal change in the level of algorithmic discrimination when moving from the conventional model to the unselected model.

To eliminate this algorithmic discrimination, we turn to our pre-processed model. The vertical axis of Panel D of Figure 1 shows the estimated $\tilde{\beta}$ coefficients from regressing true misconduct potential on the pre-processed inputs in the full sample, using estimates of the moments in Equation (7). The horizontal axis shows that, by construction, these pre-processed inputs have no conditional input disparities. Relative to the conventional model in Panel C, the pre-processed model puts more weight on having a prior misdemeanor conviction and less weight on other algorithmic inputs such as having a prior FTA and having any DUI charge.

Figure 2 also shows that eliminating algorithmic discrimination comes at little cost of accuracy. Relative to the unselected model, the pre-processed model has only a slightly higher MSE. Our three non-discriminatory algorithms also have similar accuracy, with the in-processed and post-processed models yielding similar MSEs as the pre-processed model. Taken together, these findings show that our approach offers a rare "free lunch" in this setting. Relative to the conventional model, our non-discriminatory models increase both fairness and accuracy.

Extensions and Robustness Checks. We find similar results when generating binary release recommendations at the judges' current release rates, considering different misconduct types or extrapolation approaches, or building our pre-processed model using only released defendants. For example, Appendix Figure A2 shows relatively modest levels of algorithmic discrimination at different binary release recommendations in our pre-processed model. By comparison, the conventional model recommends around a 5% lower release rate for Black defendants compared to white defendants with the same objective misconduct potential at the average detention rate in NYC. Appendix Table A3 shows that we also obtain similar results when we train our pre-processed model on alternative outcomes, including just FTA, just rearrests, or a modified outcome measure that captures the social cost of different misconduct types. We also obtain similar results when using different extrapolation methods. Finally, we show algorithmic discrimination in our pre-, in-, and post-processed models constructed using only the sample of released defendants, ignoring selection.[15] We find small and statistically insignificant levels of algorithmic discrimination, indicating that our procedures may successfully de-bias conventional algorithms even in situations in which quasi-experimental variation is not available. Whether this finding extends to other settings is an interesting question for future work.

[14] See Appendix C.1 for details on these calculations.

[15] See Appendix C.2 for details on these calculations.

4 Conclusion

Predictive algorithms often generate different predictions for protected groups, but it is unclear whether these disparities reflect algorithmic discrimination or legitimate differences in the outcome of interest. This paper develops new quasi-experimental tools to understand algorithmic discrimination and build non-discriminatory algorithms when the outcome of interest is only selectively observed, which is common in many high-stakes settings. We apply these tools in the context of pretrial bail decisions, where conventional algorithmic predictions are generated using only the misconduct outcomes of released defendants. We show that algorithmic discrimination arises when the available inputs are systematically different for white and Black defendants with the same objective misconduct potential. We then show how these conditional input disparities can be measured and purged using the quasi-random assignment of bail judges, thereby eliminating algorithmic discrimination. In NYC, our new algorithms not only eliminate algorithmic discrimination but also generate more accurate predictions by correcting for the selective observability of misconduct outcomes.

The methods developed in this paper may prove useful for understanding algorithmic discrimination and building non-discriminatory algorithms in other high-stakes settings. Our approach is appropriate whenever there is selection on the outcome of interest and experimental or quasi-experimental variation can be used to address such selection. Many settings have these features, including lending decisions, medical diagnoses, hiring decisions, and foster care decisions.

References

Albright, Alex. 2024. “The Hidden Effects of Algorithmic Recommendations.” *Unpublished Working Paper.*

Angelova, Victoria, Will S. Dobbie, and Crystal Yang. 2023. “Algorithmic Recommendations and Human Discretion.” *NBER Working Paper No. 31747.*

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. 2021. “Measuring Racial Discrimination in Algorithms.” *AEA Papers and Proceedings*, 111: 49–54.

Arnold, David, Will Dobbie, and Peter Hull. 2022. “Measuring Racial Discrimination in Bail Decisions.” *American Economic Review*, 112(9): 2992–3038.

Baron, E. Jason, Joseph J. Doyle Jr., Natalia Emanuel, Peter Hull, and Joseph Ryan. Forthcoming. “Discrimination in Multi-Phase Systems: Evidence from Child Protection.” *Quarterly Journal of Economics.*

Bergman, Peter, Elizabeth Kopko, and Julio E. Rodriguez. 2023. “A Seven-College Experiment Using Algorithms to Track Students: Impacts and Implications for Equity and Fairness.” *NBER Working Paper No. 28948.*

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *Sociological Methods & Research*, 50(1): 3–44.

Bhatt, Monica P., Sara B. Heller, Max Kapustin, Marianne Bertrand, and Christopher Blattman. 2024. “Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago.” *Quarterly Journal of Economics*, 139(1): 1–56.

Bohren, J. Aislinn, Peter Hull, and Alex Imas. 2023. “Systemic Discrimination: Theory and Measurement.” *NBER Working Paper No. 29820.*

Calmon, Flavio P., Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. “Optimized Pre-Processing for Discrimination Prevention.” *Advances in Neural Information Processing Systems*, 30: 3992–4001.

Chamberlain, Gary. 1986. “Asymptotic Efficiency in Semi-Parametric Models with Censoring.” *Journal of Econometrics*, 32(2): 189–218.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. “Selection with Variation in Diagnostic Skill: Evidence from Radiologists.” *Quarterly Journal of Economics*, 137(2): 729–783.

Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. “How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions.” *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22.

Chouldechova, Alexandra. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data*, 5(2): 153–163.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. “A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions.” *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 81: 134–148.

Coston, Amanda, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. “Counterfactual Risk Assessments, Evaluation, and Fairness.” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 582–593.

Cowgill, Bo. 2018. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening." *Columbia Business School, Columbia University.*

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–240.

Elzayn, Hadi, Evelyn Smith, Thomas Hertz, Arun Ramesh, Robin Fisher, Daniel E. Ho, and Jacob Goldin. 2023. "Measuring and Mitigating Racial Disparities in Tax Audits." *SIEPR Working Paper.*

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and Removing Disparate Impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review*, 113(1): 253–277.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *Journal of Finance*, 77(1): 5–47.

Gillette, Emma, James P. Boardman, Clara Calvert, Jeeva John, and Sarah J. Stock. 2022. "Associations Between Low Apgar Scores and Mortality by Race in the United States: A Cohort Study of 6,809,653 Infants." *PLoS Medicine*, 19(7): e1004040.

Grimon, Marie-Pascale, and Christopher Mills. 2022. "The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial." *Job Market Paper.*

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *Advances in Neural Information Processing Systems*, 29: 3323–3331.

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

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

Kallus, Nathan, and Angela Zhou. 2018. "Residual Unfairness in Fair Machine Learning from Prejudiced Data." *Proceedings of the 35th International Conference on Machine Learning*, 80: 2439–2448.

Kamiran, Faisal, and Toon Calders. 2012. "Data Preprocessing Techniques for Classification without Discrimination." *Knowledge and Information Systems*, 33: 1–33.

Kim, Michael P., Amirata Ghorbani, and James Zou. 2019. "Multiaccuracy: Black-Box Post-Processing for Fairness in Classification." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254.

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 the Fair Determination of Risk Scores." *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS).*

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.

Liang, Annie, Jay Lu, and Xiaosheng Mu. 2023. "Algorithm Design: A Fairness-Accuracy

Frontier." *Unpublished Working Paper.*

Li, Danielle, Lindsey R. Raymond, and Peter Bergman. 2020. "Hiring as Exploration." *NBER Working Paper No. 27736.*

Luminosity & The University of Chicago's Crime Lab New York. 2020. "Updating the New York City Criminal Justice Agency Release Assessment: Maintaining High Court Appearance Rates, Reducing Unnecessary Pretrial Detention, and Reducing Disparity." *New York City Criminal Justice Agency.*

Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. "Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 349–358.

Mishler, Alan, and Edward H. Kennedy. 2022. "FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1053.

Mishler, Alan, Edward H. Kennedy, and Alexandra Chouldechova. 2021. "Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 386–400.

Nguyen, Anh Thy H., Anum Saeed, Claudia E. Bambs, Justin Swanson, Nnadozie Emechebe, Fahad Mansuri, Karan Talreja, Steven E. Reis, and Kevin E. Kip. 2021. "Usefulness of the American Heart Association's Ideal Cardiovascular Health Measure to Predict Long-Term Major Adverse Cardiovascular Events (From the Heart SCORE Study)." *American Journal of Cardiology*, 138: 20–25.

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(6464): 447–453.

Ouss, Aurélie, and Megan Stevenson. 2023. "Does Cash Bail Deter Misconduct?" *American Economic Journal: Applied Economics*, 15(3): 150–182.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. "On Fairness and Calibration." *Advances in Neural Information Processing Systems*, 30: 5684–5693.

Pope, Devin G., and Justin R. Sydnor. 2011. "Implementing Anti-Discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy*, 3(3): 206–231.

Rambachan, Ashesh, Amanda Coston, and Edward H. Kennedy. 2023. "Robust Design and Evaluation of Predictive Algorithms Under Unobserved Confounding." *arXiv:2212.09844.*

Rambachan, Ashesh, and Jonathan Roth. 2020. "Bias In, Bias Out? Evaluating the Folk Wisdom." *1st Symposium on Foundations of Responsible Computing (FORC 2020), LIPIcs*, 156: 6:1–6:15.

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2023. "Algorithms, Humans and Racial Disparities in Child Protective Systems: Evidence from the Allegheny Family Screening Tool." *Unpublished Working Paper.*

Schulam, Peter, and Suchi Saria. 2017. "Reliable Decision Support Using Counterfactual Models." *Advances in Neural Information Processing Systems*, 30: 1696–1706.

Skeem, Jennifer L., and Christopher T. Lowenkamp. 2016. "Risk, Race, and Recidivism: Predictive Bias and Disparate Impact." *Criminology*, 54(4): 680–712.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review*, 103: 303–384.

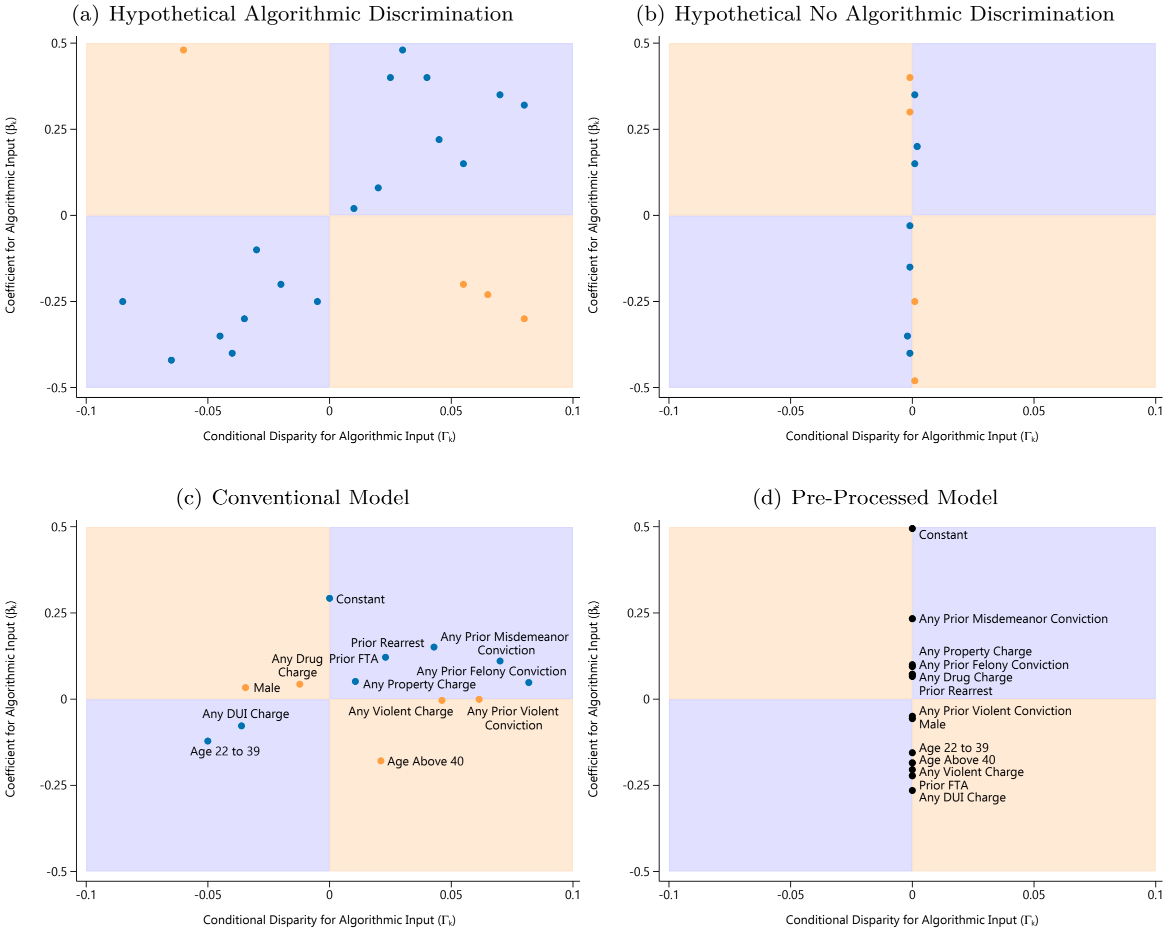
Stevenson, Megan T., and Jennifer L. Doleac. Forthcoming. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy.*

Woodworth, Blake, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. "Learning Non-Discriminatory Predictors." *Proceedings of the 2017 Conference on Learning Theory*, 65: 1920–1953.

Yang, Crystal S., and Will Dobbie. 2020. "Equal Protection Under Algorithms: A New Statistical and Legal Framework." *Michigan Law Review*, 119(2): 291–396.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment." *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180.

Figure 1: Illustrations of Algorithmic Discrimination



(a) Hypothetical Algorithmic Discrimination

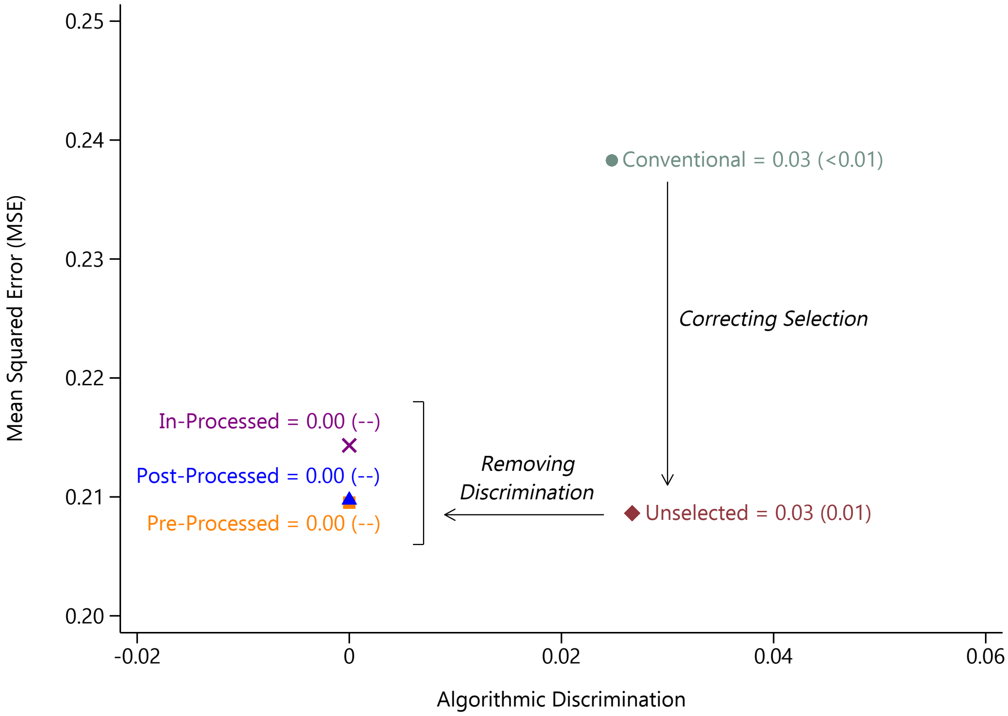
(b) Hypothetical No Algorithmic Discrimination

(c) Conventional Model

(d) Pre-Processed Model

Notes. This figure shows how conditional disparities in algorithmic inputs translate into algorithmic discrimination. The amount each point contributes to algorithmic discrimination is given by $\Gamma_k \beta_k$, where the total amount of algorithmic discrimination is $\sum_k \Gamma_k \beta_k$. Panel A illustrates a hypothetical example of algorithmic discrimination against the protected group. There are more inputs that increase algorithmic discrimination (blue) than there are inputs that decrease algorithmic discrimination (orange). Panel B illustrates a hypothetical example of no discrimination against either group. While some inputs moderately increase algorithmic discrimination (blue), others moderately decrease algorithmic discrimination (orange), leading to zero algorithmic discrimination on average. Panels C and D apply this framework to the conventional and pre-processed models, plotting model coefficients and conditional disparities for each of the 12 non-race algorithmic inputs. We estimate model coefficients using an OLS regression of an indicator for pretrial misconduct on all non-race algorithmic inputs. We measure conditional disparities using the difference in each algorithmic input between white and Black defendants with the same objective misconduct potential, computed from the unselected moments as described in the text. All points in Panel D are black as all inputs in the pre-processed model have zero conditional disparity.

Figure 2: Algorithmic Discrimination and Accuracy



Notes. This figure plots the mean squared error (MSE) of each model against algorithmic discrimination. Algorithmic discrimination captures the difference in average risk scores between white and Black defendants, conditional on true misconduct potential. Positive values indicate Black defendants on average receive higher risk scores than white defendants with the same objective misconduct potential. We report bootstrapped standard errors in parentheses. Dashes represent standard errors that are zero by construction.

Table 1: Descriptive Statistics

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Panel A: Pretrial Outcomes			
Released Before Trial	0.723	0.760	0.689
Pretrial Misconduct, When Released	0.303	0.271	0.336
Panel B: Algorithmic Inputs			
Black	0.524	0.000	1.000
Male	0.823	0.842	0.806
Age 22 to 39	0.494	0.523	0.467
Age Above 40	0.257	0.245	0.269
Prior Rearrest	0.230	0.205	0.253
Prior FTA	0.064	0.053	0.074
Any Drug Charge	0.254	0.254	0.253
Any DUI Charge	0.047	0.069	0.027
Any Violent Charge	0.148	0.130	0.165
Any Prior Felony Conviction	0.286	0.235	0.332
Any Prior Violent Conviction	0.120	0.085	0.151
Any Prior Misdemeanor Conviction	0.384	0.335	0.428
Any Property Charge	0.132	0.123	0.139
Panel C: Average Risk Scores			
Conventional Model	0.327	0.311	0.341
Total Cases	567,687	270,188	297,499
Cases with Defendant Released	410,479	205,443	205,036

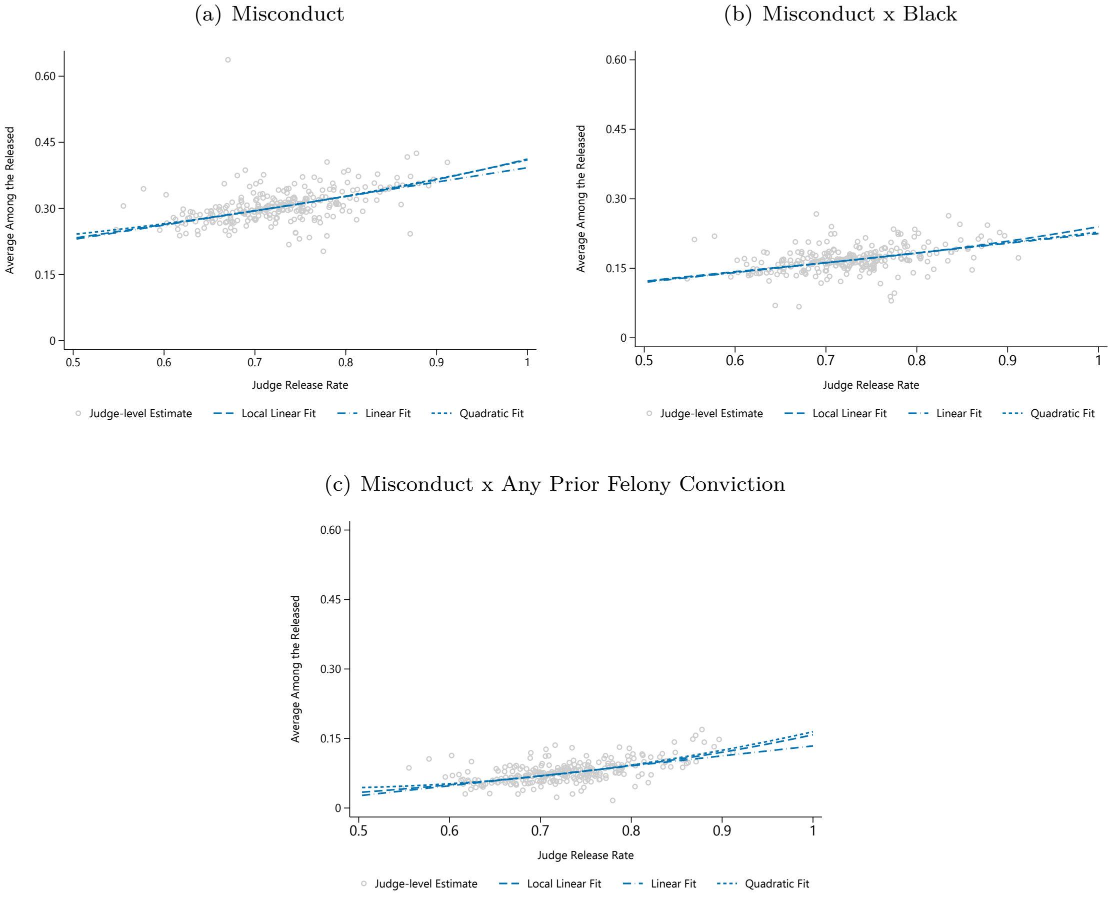
Notes. This table reports descriptive statistics for our analysis sample. The sample consists of bail hearings that were quasi-randomly assigned to judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. Pretrial misconduct is defined as either failure to appear (FTA) at a mandated court date or being rearrested for a new offense before case disposition. The conventional model regresses pretrial misconduct on the algorithmic inputs listed in Panel B, excluding race.

Appendix

Table of Contents

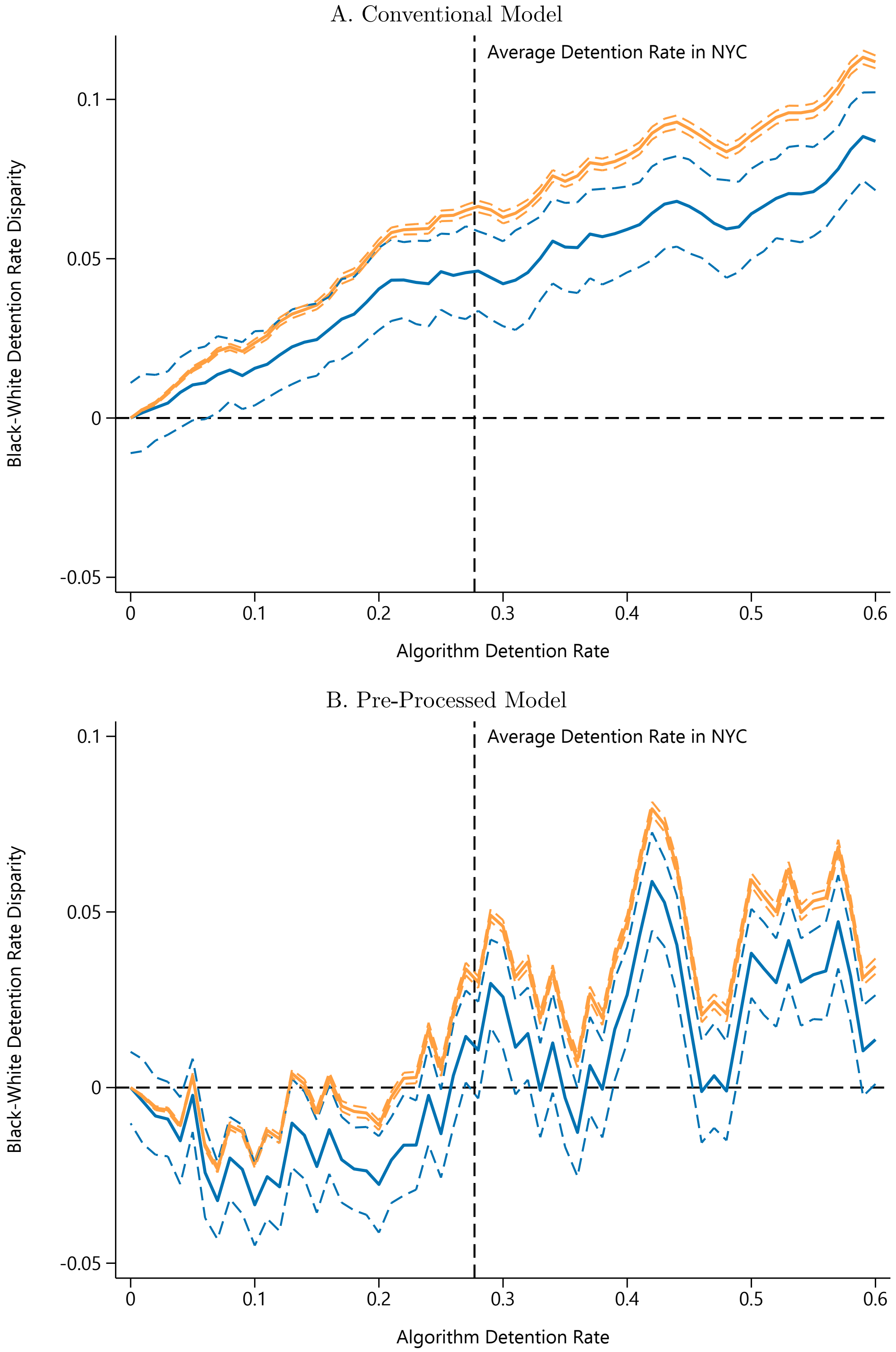
A Appendix Figures and Tables

Figure A1: Example Extrapolation Plots



Notes. This figure plots release rates for the 265 judges in our sample against the selected version of certain moments in Equation (7). Panel A extrapolates pretrial misconduct, our main outcome of interest, to estimate $E[Y_i^*]$. Panel B extrapolates the interaction of pretrial misconduct and race to estimate $E[G_i Y_i^*]$. Panel C extrapolates the interaction of pretrial misconduct and any prior felony conviction to estimate $E[X_{ik} Y_i^*]$, where k in this example refers to any prior felony conviction. All estimates adjust for court-by-time fixed effects as described in Equations (11) and (12). The dotted curves plot the fitted values of local linear, linear, and quadratic regressions that inverse-weight by the variance of the estimated released second moments.

Figure A2: Unconditional Disparities and Algorithmic Discrimination in Detention Rates


A. Conventional Model

B. Pre-Processed Model

Notes. This figure plots unconditional disparities and algorithmic discrimination in detention rates for various thresholds of detention. In Panel A, we first estimate a conventional model and then rank defendants by predicted risk of misconduct. The unconditional disparity is the difference in detention rates for Black defendants relative to white defendants for a given detention rate in the population. Algorithmic discrimination is defined as the difference in detention rates for Black defendants relative to white defendants, conditional on the same objective misconduct potential. Panel B uses the pre-processed model to rank defendants by predicted risk of misconduct. Dashed lines indicate pointwise 95 percent confidence intervals obtained from a bootstrapping procedure.

Table A1: Tests of Quasi-Random Judge Assignment

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Male	0.00008	0.00010	0.00007
	(0.00012)	(0.00016)	(0.00016)
Age 22 to 39	-0.00013	-0.00020	-0.00006
	(0.00012)	(0.00017)	(0.00017)
Age Above 40	-0.00022	-0.00028	-0.00015
	(0.00012)	(0.00019)	(0.00016)
Prior Rearrest	-0.00003	0.00015	-0.00018
	(0.00009)	(0.00015)	(0.00012)
Prior FTA	-0.00027	-0.00018	-0.00032
	(0.00019)	(0.00026)	(0.00025)
Any Drug Charge	-0.00007	-0.00008	-0.00008
	(0.00013)	(0.00018)	(0.00016)
Any DUI Charge	0.00041	0.00046	0.00014
	(0.00023)	(0.00026)	(0.00037)
Any Violent Charge	0.00008	-0.00013	0.00022
	(0.00018)	(0.00026)	(0.00019)
Any Prior Felony Conviction	-0.00022	0.00001	-0.00038
	(0.00013)	(0.00020)	(0.00015)
Any Prior Violent Conviction	-0.00015	-0.00027	-0.00007
	(0.00015)	(0.00020)	(0.00019)
Any Prior Misdemeanor Conviction	0.00018	0.00013	0.00020
	(0.00011)	(0.00014)	(0.00014)
Any Property Charge	-0.00029	-0.00029	-0.00030
	(0.00015)	(0.00017)	(0.00023)
Black	-0.00011		
	(0.00008)		
Joint p-value	[0.12304]	[0.44736]	[0.09443]
Court x Time FE	Yes	Yes	Yes
Cases	567,687	270,188	297,499

Notes. This table reports OLS estimates of regressions of judge leniency on various defendant and case 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, following the procedure in Arnold, Dobbie and Hull (2021). All regressions control for court-by-time fixed effects. The p-values reported at the bottom of each column are from F-tests for 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.

Table A2: First-Stage Effects of Judge Leniency

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Judge Leniency	1.169	0.968	1.340
	(0.020)	(0.028)	(0.033)
Court x Time FE	Yes	Yes	Yes
Mean Release Rate	0.723	0.760	0.689
Cases	567,687	270,188	297,499

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, following the procedure in Arnold, Dobbie and Hull (2021). 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.

Table A3: Extensions and Robustness Checks

		Alternative Outcomes			Alternative Extrapolations		
	Baseline	Failure to Appear	Any Rearrest	Social Cost	Linear	Quadratic	Released Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Conventional	0.025	0.011	0.024	243.145	0.024	0.027	0.025
	(0.003)	(0.001)	(0.002)	(20.728)	(0.001)	(0.003)	(0.003)
Unselected	0.027	0.012	0.027	121.354	0.021	0.042	0.025
	(0.009)	(0.006)	(0.007)	(145.806)	(0.004)	(0.013)	(0.003)
Pre-Processed	0.000	0.000	0.000	0.000	0.000	0.000	0.002
	–	–	–	–	–	–	(0.004)
In-Processed	0.000	0.000	0.000	0.000	0.000	0.000	0.004
	–	–	–	–	–	–	(0.002)
Post-Processed	0.000	0.000	0.000	0.000	0.000	0.000	0.003
	–	–	–	–	–	–	(0.003)
Cases	567,687	567,687	567,687	567,687	567,687	567,687	567,687

Notes. This table shows the main results from our baseline specification along with various extensions and robustness checks. Column 1 reports algorithmic discrimination in pretrial misconduct, our main outcome of interest, where algorithmic discrimination is the coefficient from regressing predicted pretrial misconduct on a Black indicator, controlling for true misconduct potential. Columns 2, 3, and 4 report algorithmic discrimination when using alternative outcomes: failure to appear, any rearrest, and the social cost of misconduct. We construct our social cost measure as described in Angelova, Dobbie and Yang (2023). Column 5 reports algorithmic discrimination when using a simple linear extrapolation method, relative to our baseline local linear approach. Column 6 reports algorithmic discrimination when using a quadratic extrapolation method. Column 7 reports algorithmic discrimination when selection correction is done using only the sample of released defendants. We report bootstrapped standard errors in parentheses. Dashes represent standard errors that are zero by construction. See Table 1 for details on the sample.

B Data and Setting Appendix

B.1 Judge Assignment and Decisions in NYC

In the main text, we describe judges as making a binary release or detain decision. However, in practice, judges generally have a few options. They can choose to release-on-recognizance (ROR), in which case the defendant is released without conditions. The judge can also charge a monetary bail that a defendant must pay in order to be released. The bail amount is up to the judge's discretion. The bail amount will be returned if the defendant appears at all future mandated court dates. A defendant may also go through a bail bondsman, who will post bail for a fee. If the defendant is unable or unwilling to pay this fee, then the defendant will remain detained until trial. Finally, a judge can deny the possibility of bail altogether and remand the defendant into custody. Misconduct outcomes are unobserved for both defendants that are remanded and defendants that are unable to pay their monetary bail.

During a case, a judge is presented with a variety of information about the defendant, including details of the arrest and charge. Since 2003, judges have also been given a risk assessment tool that predicts whether a defendant would fail to appear in court. This risk assessment was updated in November 2019 (Luminosity & The University of Chicago's Crime Lab New York, 2020). While the two risk assessments vary in terms of the algorithmic inputs and weights associated with them, both are linear in a small number of characteristics.

Cases are assigned to judges in NYC using a rotation calendar system in each of the five county courthouses, generating quasi-random variation in bail judge assignment for defendants arrested at the same time and in the same place. Each county courthouse employs a supervising judge to determine the schedule that assigns bail judges to the day (9 a.m. to 5 p.m.) and night arraignment shifts (5 p.m. to 1 a.m.) in one or more courtrooms within each courthouse. Individual judges can request to work certain days or shifts but in practice, there is considerable variation in judge assignments within a given arraignment shift, day-of-week, month, and year cell. Our assumption is that within these court-by-time cells (i.e., assigned courtroom, shift, day-of-week, month, and year cells), the judge assigned to a given defendant is randomly selected.

To test this assumption, Appendix Table A1 reports coefficients from an ordinary least squares (OLS) regression of judge leniency on various defendant and case characteristics, controlling for court-by-time fixed effects. We measure leniency using the leave-one-out average release rate among all other defendants assigned to a defendant's judge.[1] Most coefficients in this balance table are small and not statistically significantly different from zero, both overall and by defendant race. A joint F-test fails to reject the null of quasi-random assignment at conventional levels of statistical significance.

Appendix Table A2 verifies that judge assignment meaningfully affects the probability that a defendant is released before trial. Each column of this table reports coefficients from an OLS regression of an indicator for pretrial release on judge leniency and court-by-time fixed effects. A one percentage point increase in the predicted leniency of a defendant's judge is associated with a 1.17 percentage point increase in the probability of release, with a somewhat smaller first-stage effect for white defendants

[1] Following the standard approach in the literature (e.g., Arnold, Dobbie and Yang, 2018; Dobbie, Goldin and Yang, 2018; Arnold, Dobbie and Hull, 2022), we construct the leave-one-out measure by first regressing pretrial release on court-by-time fixed effects and then using the residuals from this regression to construct the leave-one-out residualized release rate. By first residualizing on court-by-time effects, the leave-one-out measure captures the leniency of a particular judge relative to that of judges assigned to the same court-by-time cells.

and a somewhat larger effect for Black defendants.

B.2 Sample-Selection Criteria

We make six key restrictions to arrive at our estimation sample, broadly following Arnold, Dobbie and Hull (2022). First, we drop cases where the defendant is not charged with a felony or misdemeanor (N=26,057). Second, we drop cases that were disposed at arraignment (N=364,051) or adjourned in contemplation of dismissal (N=230,517). Third, we drop cases in which the defendant is assigned a \$1 cash bail (N=1,284). Cash bail is assigned if the defendant is already serving time in jail on an unrelated charge. The \$1 cash bail is set so that the defendant receives credit for served time and does not reflect a new judge decision. Fourth, we drop defendants who are non-white and non-Black (N=45,529). Fifth, we drop cases for which a defendant received a desk appearance ticket since a desk appearance ticket does not require an arraignment hearing (N=76,232). Finally, we drop defendants assigned to judges with fewer than 100 cases (N=3,637) and court-by-time cells with fewer than 100 cases or only one unique judge (N=143,062), where a court-by-time cell is defined by the assigned courtroom, shift, day-of-week, month and year (e.g., the Wednesday night shift in Courtroom A of the Kings County courthouse in January 2012). The final sample consists of 567,687 cases, 353,422 defendants, and 265 judges. Relative to the full sample of cases, our estimation sample has a somewhat lower release rate, although the ratio of release rates by race is similar. Our estimation sample is also broadly representative in terms of defendant and charge characteristics, with a slightly lower share of defendants with prior FTAs, a slightly higher share of defendants with prior rearrests, and a lower share of defendants charged with drug and property crimes.

C Econometric Appendix

C.1 Mean Squared Error Calculation

The mean squared error (MSE) of the predictions of a linear algorithm $\hat{Y}_i = X_i'\beta$ is given by:

$$E[(Y_i^* - \hat{Y}_i)^2] = E[Y_i^{*2}] - 2E[X_i'Y_i^*]\beta + \beta' E[X_i X_i']\beta, \tag{C1}$$

where the first and second terms are elements of Θ. These are affected by the selective observability of Y_i^* and are estimated as part of our main analysis. The third term is not affected by selective observability of Y_i^* and is directly estimable. We use this formula to compute the MSE of the conventional, unselected , and in-processed models.

The MSE of the pre-processed model predictions $\hat{Y}_i^{Pre} = \tilde{X}_i'\tilde{\beta}$, where $\tilde{X}_i = X_i - \Gamma G_i$, is given by:

$$\begin{aligned} E[(Y_i^* - \hat{Y}_i^{Pre})^2] =& E[Y_i^{*2}] - 2E[\tilde{X}_i'Y_i^*]\tilde{\beta} + \tilde{\beta}' E[\tilde{X}_i\tilde{X}_i']\tilde{\beta} \\ =& E[Y_i^{*2}] - 2E[X_i'Y_i^*]\tilde{\beta} - 2E[G_iY_i^*]\Gamma'\tilde{\beta} \\ &+ \tilde{\beta}'E[X_iX_i']\tilde{\beta} - \tilde{\beta}'\Gamma E[G_iX_i']\tilde{\beta} - \tilde{\beta}'E[X_iG_i]\Gamma'\tilde{\beta} + \tilde{\beta}'\Gamma E[G_i]\Gamma'\tilde{\beta}, \end{aligned} \tag{C2}$$

which is again a function of the elements of Θ (both directly estimable and selection-affected) and the directly estimable $E[X_iX_i']$. Finally, the MSE of the post-processed model predictions $\hat{Y}_i^{Post} =$

$X_i'\beta^U - \Delta^U G_i$ is given by:

$$\begin{aligned} E[(Y_i^* - \hat{Y}_i^{Post})^2] =& E[(Y_i^* - X_i'\beta^U)^2] + 2\Delta E[G_i(Y_i^* - X_i'\beta^U)] + \Delta^2 E[G_i] \\ =& E[Y_i^{*2}] - E[X_i'Y_i^*]\beta^U + 2\Delta\left(E[G_iY_i^*] - E[X_i'G_i]\beta^U\right) + \Delta^2 E[G_i], \end{aligned} \tag{C3}$$

where we use the fact that $\beta^U = E[X_iX_i']^{-1}E[X_iY_i^*]$ to simplify in the second line. This is also a function of the elements of Θ (both directly estimable and selection-affected).

C.2 Algorithmic Discrimination in Released-Sample Models

Column 7 of Appendix Table A3 shows estimates of algorithmic discrimination for versions of the pre-processed, in-processed, and post-processed models that use released-sample observations of Y_i^* to adjust the conventional model instead of our baseline quasi-experimental selection-correction approach. We detail these calculations below.

Released-sample pre-processed model predictions are given by $\tilde{X}_i^{R\prime}\beta^R$, where $\tilde{X}_{ik}^R = X_{ik} - \Gamma_k^R G_i$. Γ_k^R is the coefficient on G_i from running regression (3) in the $D_i = 1$ subpopulation. The β^R coefficient vector is obtained from regressing Y_i^* on the set of $\tilde{X}_{ik}^R$ in the $D_i = 1$ subpopulation. The level of algorithmic discrimination in this model is given by $(\Gamma - \Gamma^R)'\beta^R$, where Γ collects the coefficients on G_i from the full-population regression (3), estimated using quasi-experimental variation.

Released-sample in-processed model predictions are given by $X_i'\beta^{*R}$, where β^{*R} is given by a released-sample version of Equation (6):

$$\beta^{*R} = \beta - E[X_iX_i']^{-1}\Gamma^R\left(\Gamma^{R\prime}E[X_iX_i']^{-1}\Gamma^R\right)^{-1}\Gamma^{R\prime}\beta, \tag{C4}$$

which replaces (β^U, Γ) with (β, Γ^R). β is the coefficient from regressing Y_i^* on the set of X_{ik} in the $D_i = 1$ subpopulation. The level of algorithmic discrimination in this model is $\Gamma'\beta^{*R}$.

Released-sample post-processed model predictions are given by $X_i'\beta - \Delta^R G_i$, where $\Delta^R = \Gamma^R\beta$ is a released-sample measure of algorithmic discrimination for the conventional model, i.e., the coefficient from regressing $X_i'\beta$ on G_i controlling for Y_i^* in the $D_i = 1$ subpopulation. The level of algorithmic discrimination in this model is given by $\Gamma'\beta - \Delta^R$.