Reconciling Legal and Empirical Conceptions of Disparate Impact

Joshua Grossman
Stanford University

Julian Nyarko
Stanford University

Sharad Goel
Harvard University

Abstract

In this paper, we consider the statistical foundations for empirical tests of disparate impact. We begin by considering a recent, popular proposal in the economics literature that seeks to assess disparate impact via a comparison of error rates for the majority and the minority group. We show that this approach suffers from what is colloquially known as "the problem of inframarginality", in turn putting it in direct conflict with legal understandings of discrimination. We then proceed to analyze two alternative proposals that quantify disparate impact either in terms of risk-adjusted disparities or by comparing existing disparities to those under a statistically optimized decision policy. Both approaches have differing, context-specific strengths and weaknesses, and we discuss how they relate to the individual elements in the legal test for disparate impact. To demonstrate feasibility, we assess disparate impact in a large dataset of 2.2 million pedestrian stop-and-frisk decisions recorded by the New York City Police Department between 2008 and 2011. We find strong evidence for disparate impact and propose both complex and simple policy alternatives that are as efficient while exerting fewer disparities.

1 Introduction

Anti-discrimination law in the U.S. recognizes two distinct forms of discriminatory conduct. First, disparate treatment law aims at prohibiting decisions that are fueled by discriminatory motivations. This notion of discriminatory conduct is largely consistent with Becker's popular model of discrimination, which defines as discriminatory those actions that are either motivated by animus ("taste-based discrimination") or that use race as a proxy for an unobservable, decision-relevant factor ("statistical discrimination") [Becker, 1957]. But in many areas of life, such as employment and credit, U.S. law has long embraced a second definition of discrimination, significantly broadening its scope. This notion of discrimination is known as disparate *impact*. The doctrine of disparate impact renders illegal those policies that produce avoidable and unjustified excess disparities [Griggs v. Duke Power Co., 401 U.S. 424 (1971)]. In practice, disparate impact is often found if the plaintiff can demonstrate the existence of an alternative, feasible decision rule that is at least as good as the existing decision rule at achieving the stated policy goal while imposing fewer disparities. Although the law's embrace of disparate impact doctrine can be traced back decades, empirical scholarship both in the law and in the social sciences at large has almost exclusively limited itself to the analysis of disparate treatment. It is only recently that the

literature has made a systematic attempt to broaden its focus [Arnold et al., 2022, Ayres, 2005, 2010, Bohren et al., 2022, Cai et al., 2022, Elzayn et al., 2023, Grossman et al., 2023, Grunwald et al., 2022, Jung et al., 2019]. In its wake, several statistical frameworks for the measurement of disparate impact have been proposed independently.

In this paper, we introduce, compare and critically assess the three most prominent recent proposals: risk-adjusted disparities, disparities relative to statistically optimized decision policies, and error rate disparities. We discuss their individual advantages and disadvantages, and examine how they relate to the legal doctrine of disparate impact as it has been developed by U.S. courts. We argue that the first two proposals speak to different legal elements of disparate impact doctrine, making both valuable empirical tools for assessing disparate impact in various situations, albeit with context-specific strengths and weaknesses. However, we further argue that the third proposal—error-rate disparities—is generally unsuitable for assessing disparate impact. To foreshadow our argument, assume a judge who makes detention decisions by balancing public safety with the individual rights of the defendants. The judge is able to perfectly distinguish between risky and non-risky defendants, and chooses to only detain risky defendants. Assuming further that detention decisions were generally subject to disparate impact law,[1] the judge's decision practice would nonetheless not be considered to exert a disparate impact. After all, the judge is making decisions that optimally fulfill the goal of balancing public safety with the defendant's interests, and any residual disparities that result would thus be considered justified. Yet, as we show below, in most scenarios a measure that relies on error rates would find that the judge's decision practice is illegal, a result that can be reconciled neither with disparate impact doctrine nor with existing normative notions of discrimination. At a technical level, we show that error-rate-based measures suffer from what is known as the "problem of inframarginality" [Ayres, 2002, Simoiu et al., 2017].

Finally, we utilize the insights obtained from our discussion and apply them to the concrete example of stop-and-frisk decisions. We analyze a dataset of all 2.2 million pedestrian stops recorded by the New York City Police Department between 2008 and 2011. During this time period, officers could choose to frisk stopped individuals whom they perceived as sufficiently likely to be carrying a weapon. Officers were, on average, more likely to frisk stopped Black individuals than stopped white individuals, providing *prima facie* evidence of disparate impact.

Moving beyond this *prima facie* evidence, we then apply the first two proposals above to assess the evidence for disparate impact. Following the first proposal, we compute risk-adjusted disparities to determine whether the gap in frisk rates is justified by legitimate policy goals—namely, the recovery of weapons. To do so, we estimate the statistical likelihood that an individual is carrying a weapon, using all available recorded information. We find that frisk rates for stopped Black individuals are considerably larger than for comparably risky stopped white individuals (i.e., the racial disparities persist even after accounting for risk). Next, following the second proposal, we compare the observed racial disparities to those achievable under a set of statistically optimized alternative frisk policies. We specifically consider a set of "threshold" policies, in which all individuals above a given level of estimated risk are frisked. We find that there are indeed alternative policies that: (1)

[1] Disparate impact is only illegal if a statute deems it so.

recover more weapons than the status quo; (2) require conducting fewer frisks; and (3) have smaller racial disparities. The existence of such policies provides further evidence of disparate impact.

Embracing a broader concept of anti-discrimination is vital in ensuring that empirical scholarship remains closely tied to legal realities. Our hope is that this study can contribute to that goal by serving as a guide to researchers interested in assessing the disparate impact of a policy or decision rule.

2 The Law of Disparate Treatment and Disparate Impact

Although our focus lies on disparate impact law, we believe a brief primer on the legal concepts in U.S. anti-discrimination can serve as helpful background. For ease of exposition, we will focus on the law surrounding racial discrimination, although most of the content equally applies to other forms of discrimination based on protected features, such as gender.

Generally speaking, U.S. law recognizes two forms of discriminatory conduct: disparate treatment and disparate impact. Disparate treatment encapsulates the most intuitive notion of discrimination. It is aimed at outlawing decisions and policies that are motivated by race, making discriminatory intent the crucial element of disparate treatment [DeJung v. Superior Ct., 169 Cal. App. 4th 533 (2008); McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973)]. The intent can take the form of explicit, racially conditioned decision making. But more commonly, disputes focus on facially neutral decisions or policies that are alleged to be—at least in part—racially motivated. If discriminatory intent is present and the discriminatory actor is a public entity or official, the decision is subjected to judicial review under a "strict scrutiny" standard [United States v. Carolene Prod. Co., 304 U.S. 144 (1938)].[2] This standard is very difficult to meet and requires that the conduct in question is *narrowly tailored* to serve a *compelling state interest*. The only cases relevant today in which race-based decisions met this standard consist of affirmative action cases in a handful of domains, such as in government contracting [Rothe Dev., Inc. v. United States Dep't of Def., 836 F.3d 57 (D.C. Cir. 2016)] and—until recently—education [Fisher v. Univ. of Texas at Austin, 579 U.S. 365 (2016); Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll., 143 S. Ct. 2141 (2023)]. Disparate treatment by public entities is always illegal under the Equal Protection Clause of the U.S. Constitution. In addition, disparate treatment of private actors is outlawed by federal and state laws in many public-facing contexts,[3] although relevant nuances may vary by context.[4]

Due to the strong emphasis on the element of discriminatory intent, it is common to conceptualize disparate treatment as decisions made *because of* race.[5] Empirical legal research

[2]The now widely disparated case of Korematsu, although now overturned, is also among the cases that contributed to the development of the strict scrutiny standard. Korematsu v. United States, 323 U.S. 214 (1944).

[3]A notable exception is insurance, where disparate treatment is not outlawed in all states [Avraham et al., 2013].

[4]For instance, 29 C.F.R. § 1608 lays out detailed guidelines under which voluntary affirmative action efforts by private employers are protected.

[5]The case law has since developed to also include other notions of disparate treatment. For instance, even if a policy was originally instituted with good intentions, under the concept of "deliberate indifference", some courts have found it can still constitute disparate treatment if the policy's disproportionate, negative

has translated this definition of disparate treatment into a "kitchen-sink" model [Gaebler et al., 2022]. Under this approach, the investigator typically runs a regression of the form

$$\Pr(\mathtt{Y}_i = 1) = \text{logit}^{-1}\left(\alpha_{\mathtt{race}[i]} + \beta^T \vec{X}_i\right), \tag{1}$$

with $\mathtt{Y}_i$ representing a binary outcome of the decision, $\alpha_{\mathtt{race}}$ an intercept term shared by defendants with the same race or ethnicity as defendant i, and $\vec{X}$ is a vector of additional controls. The controls included in $\vec{X}$ are typically large. The idea behind the kitchen-sink approach is that $\vec{X}$ controls for non-racial factors that might motivate the decision (e.g., to detain). Thus, any residual variation that is explained by $\alpha_{\mathtt{race}}$ holding the covariates in $\vec{X}$ constant is taken as evidence of discriminatory intent. Following this logic, Wooldredge [2012] seeks to provide evidence for disparate treatment of Black pretrial defendants by fitting models of the form above that adjust for legally relevant controls, including demographics, prior criminal history, and charges. Many other studies, especially in criminal law, follow a similar process [Bridges and Steen, 1998, Demuth, 2003, Didwania, 2020, Donnelly and MacDonald, 2018, Metcalfe and Chiricos, 2018, Rehavi and Starr, 2014].

There are some problems with conceiving of disparate treatment in this way. Among others, it is our view that empirical researchers often define the set of covariates included in $\vec{X}$ too broadly. Because every variable in $\vec{X}$ is implicitly accepted as being free of racial motivation, being too broad can quickly lead researchers to mask discriminatory intent if the discriminatory practice is implemented through a facially neutral factor. However, a full discussion of statistical measures of disparate treatment is beyond the scope of this paper.

In addition to disparate treatment, U.S. anti-discrimination laws sometimes render illegal a second form of discriminatory conduct, disparate impact. But unlike disparate treatment, there is no general prohibition of disparate impact under the U.S. Constitution. Instead, disparate impact is rendered illegal only through state and federal laws. The most prominent domains subject to disparate impact analysis include credit [15 U.S.C. § 1691 et seq.], employment [42 U.S.C. § 2000e et seq.] and housing [42 U.S.C. § 3601 et seq.]. Although the fragmented nature requires a few generalizations, disparate impact laws aim to prevent policies and decisions that, while not necessarily racially motivated, nonetheless have an adverse impact on racial minorities that cannot be justified by a furtherance of the policy goals.

To illustrate, consider the case of a job posting by a tech company for the position of a software engineer. The posting requires applicants to have a computer science degree. The degree requirement impacts Black potential applicants more negatively than white potential applicants, given that the share of Black computer scientists is disproportionately low [Dillon Jr et al., 2015]. However, a computer science degree can reasonably be assumed to teach skills that software engineers benefit from, meaning that the degree requirement does not constitute disparate impact. But contrast this to the seminal case of *Griggs v. Duke Power Co.*, where the Supreme Court examined an internal policy under which a high school diploma was required for certain promotions within Duke Power Company in North

impact on minorities is known and the policy is not corrected within a reasonable time frame [Davis Next Friend LaShonda D. v. Monroe Cnty. Bd. of Educ., 526 U.S. 629 (1999); Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013)].

Carolina. Black employees were much less likely to hold a high school diploma than white employees, thus disproportionately excluding the Black minority from the positions. The Supreme Court found that, while it is principally permitted to impose job requirements that impact racial minorities disproportionately, a high school diploma did not indicate better job performance, thus rendering the requirement illegal.

More formally, legal tests of disparate impact typically have three elements. Those require that: (1) the minority group is disproportionately impacted by a policy ("adverse impact") [New York City Env't Just. All. v. Giuliani, 214 F.3d 65 (2d Cir. 2000)]; (2) that there is no legitimate justification for the policy [Texas Dep't of Hous. & Cmty. Affs. v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015)]; and (3) that an alternative policy with a lesser disproportionate impact is available and implementable [Elston v. Talladega Cnty. Bd. of Educ., 997 F.2d 1394 (11th Cir. 1993)]. The plaintiff is responsible for establishing that the defendant's policy adversely impacts the minority group. The burden then shifts to the defendant, who must show that the adverse impact is justified by legitimate policy goals. Failing to do so would typically result in a finding of disparate impact. However, if the defendant does provide compelling justification for the disparities, the burden then shifts back to the plaintiff, who, to establish a finding of disparate impact, must show that there exists an equally efficient policy with less adverse impact than the status quo [see, e.g., 42 U.S. Code § 2000e–2]. We describe these three elements in more detail below.

Adverse Impact An adverse impact is typically defined as the difference in group-based selection rates [29 C.F.R § 1607.16]. In the context of standardized tests for promotions, for instance, a court would compare the passage rate among white test takers to the passage rate among Black test takers. The test would demonstrate an adverse impact if the passage rate among Black test takers was substantially lower than that among white test takers.[6]

No Justification An adverse impact lacks a substantial justification if it is not demonstrably related to a significant, legitimate goal. At times, it is also held that the adverse impact needs to be a *necessary condition* to effectuate the policy goal.[7] How courts operationalize this requirement is highly context-specific [Clady v. Los Angeles Cnty., 770 F.2d 1421 (9th Cir. 1985); Smith v. Xerox Corp., 196 F.3d 358, 363 (2d Cir. 1999); Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518 (M.D. Ala. 1991)]. For instance, the strength of the evidence required may vary by the extent of the adverse impact, by the entity that makes the relevant decision, and by whether the decision-relevant factors that cause the disparity are innate or can be acquired.

No Less Discriminatory Alternative Demonstrating the shortcomings of the current policy is not enough if there is no less discriminatory alternative [Elston v. Talladega

[6] In the employment context, courts often apply a four-fifths rule, under which the difference is significant if the passage rate for Black test takers is less than 80% of the passage rate of white test takers [29 C.F.R. § 1607.4].

[7] In which case the dividing line between the justification requirement and the requirement for a less discriminatory alternative is blurred.

Cnty. Bd. of Educ., 997 F.2d 1394 (11th Cir. 1993); Georgia State Conf. of Branches of NAACP v. State of Ga., 775 F.2d 1403 (11th Cir. 1985)]. In this way, disparate impact law is grounded within the realm of feasible policy choices: If the only way for an employer to mitigate adverse impact is to spend tens of thousands of dollars on each applicant to assess their suitability for the job, this is not something that anti-discrimination laws will ask of them. With the advent of algorithmic decision making, the requirement to have no less discriminatory alternative has received heightened relevance. Often, if a decision was based on these complex model estimates, it would both improve outcomes and decrease the adverse impact [Goel et al., 2016]. However, it remains unclear in what contexts decision makers will be required to rely on these more complex estimation procedures. Does disparate impact law require employers to forego their traditional, interview-based hiring practices if it can be shown that algorithmic assessments of job performance are superior and impose fewer disparities [Hoffman et al., 2018]? To date, courts have shied away from providing a clear answer.

3 Statistical formulations of disparate impact

Unlike for disparate treatment, there have been surprisingly few attempts to provide a statistical framework for the evaluation of disparate impact. Our goal in this section is to introduce and mediate between the different approaches. We focus on three statistical formulations, all of which are relatively recent.

3.1 Differences in error rates

One approach to measuring disparate impact is rooted in error rates. This approach deems discriminatory those decisions that lead to differences in error rates across the marginalized and the majority group, such as the false positive or the false negative rate. Conceiving of biases as error rates has a long tradition in the literature on algorithmic fairness in computer science and statistics [Buolamwini and Gebru, 2018, Chouldechova, 2017, Corbett-Davies et al., 2017, Dwork et al., 2012, Kleinberg et al., 2017], law [Chander, 2016, Huq, 2019, Mayson, 2019], medicine [Goodman et al., 2018, McCradden et al., 2020, Paulus and Kent, 2020], the social sciences [Berk et al., 2021, Imai et al., 2023, Kleinberg et al., 2018], and philosophy [Card and Smith, 2020, Hu and Kohler-Hausmann, 2020, Kasy and Abebe, 2021]. However, this formulation of bias has, for a long time, not been directly tied to disparate impact. But a recent contribution in the economics literature has proposed a measure based on error rates that is explicitly described as an estimand corresponding to the legal concept of disparate impact [Arnold et al., 2021, 2022, Baron et al., 2023]. This estimand—and its associated, novel estimation method—have since attracted significant attention.

Arnold et al. [2022] illustrate their measure of disparate impact in a pretrial detention setting in which judges must decide whether or not to release defendants on bail. Each defendant has a latent misconduct potential, which takes on the value 1 if the defendant will violate the terms of release if released, and 0 if not. Their measure of disparate impact, Δ, is based on a weighted sum of the difference in true negative rates and the difference in false negative rates across two groups of individuals, with weights defined by the overall violation rate across all individuals. Arnold et al. [2022] use the following mathematical

formulation:

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

where δ_r^T is the true negative rate for individuals of race r (i.e., among those released, the proportion that will not violate), δ_r^F the false negative rate for individuals of race r (i.e., among those released, the proportion that does violate), and $\bar{\mu}$ the expected violation rate if all individuals were released. Here, for exposition, w refers to white defendants, and b refers to Black defendants.

3.1.1 The problem with error rates

We believe that any measure of disparate impact (or fairness, for that matter) that is based on error rates is ill-suited to provide either legal or policy guidance. This is because these measures suffer from what is colloquially known as the "problem of inframarginality" [Ayres, 2002]. Intuitively, the problem is that error rates do not only capture aspects of the decision rule, but also of the underlying risk distribution for each group [Simoiu et al., 2017]. When defining disparate impact in such a way, an actor who does the best possible job to make decisions in furtherance of the stated policy goal can be found to discriminate simply due to differences in underlying risk distributions. In an attempt to avoid liability for disparate impact under this definition, the actor would then be required to make contra-indicated decisions, such as to frisk or jail people that, to the best of their knowledge, are of low risk.

To illustrate by way of a specific example in the context using the Δ estimator proposed by Arnold et al. [2022], consider Figure 1. Suppose there are two groups of pretrial defendants, each with 100 defendants. Each defendant has either a 10% likelihood of violating the terms of pretrial release if released ("low risk") or a 40% likelihood ("high risk").[8] 30% of defendants in Group 1 are of high risk, and 40% of defendants in Group 2 are of high risk. Further suppose that the presiding judge can perfectly perceive whether a defendants is of low risk or high risk. The judge decides whether to detain defendants based on a simple rule: high risk defendant are detained, and low risk defendants are released. Denote a true negative as an instance in which the judge releases a low-risk defendant. Denote a false negative as an instance in which the judge releases a high-risk defendant.

Although this decision rule treats similarly situated[9] defendants identically, the true negative rates and false negative rates among each group differ in expectation. In this example, the true negative rate is the proportion who are released, among those who *would not* violate if released. Here, 81 of the defendants in Group 1 would not violate if released (as indicated by the circles in the left-hand side of Figure 1). Further, 63 of these defendants are actually released—represented by the ◯ symbols above the dotted line—resulting in a true negative rate of $63/81 = 78\%$. We can analogously compute the true negative rate for Group 2. In particular, among the 78 defendants from Group 2 who would not violate if released (the ◯ symbols on the right-hand side of Figure 1), 54 are released (those above the

[8] For simplicity, we use groups of equal size with only two possible risk levels. This particular example is amenable to groups of different size, with two unique risk levels for each group. In Appendix Figure A1, we extend the example to a setting with a continuous distribution of risk.

[9] Where similarly situated is with respect to the goal of the release policy, which is to release as many defendants as possible while minimizing pretrial violations.

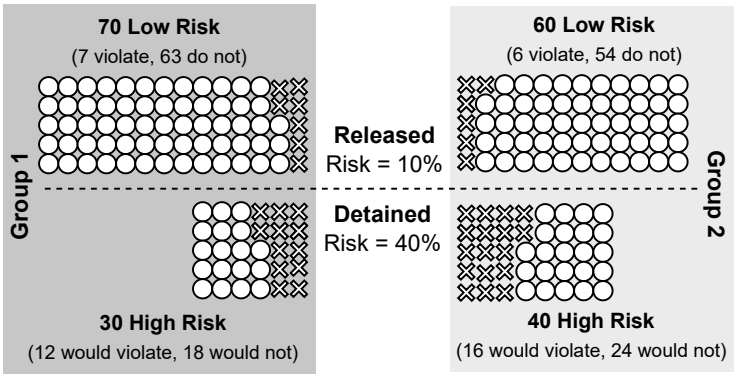

Figure 1: *Illustration of the problem of inframarginality when comparing error rates across groups with different underlying distributions of risk. Suppose there are two groups of pretrial defendants and two possible levels of pretrial risk. Each group has 100 defendants. If released pretrial, lower risk defendants violate the terms of release 10% of the time, and higher risk defendants violate 40% of the time. 30% of Group 1 defendants are higher risk, compared to 40% of Group 2 defendants. Suppose a judge can perfectly perceive pretrial risk. The judge imposes a unilateral risk threshold decision rule in deciding whom to release: lower risk defendants are always released, and higher risk defendants are always detained. In this scenario, the* Δ *measure of disparity from Arnold et al. [2022] is approximately* 0.1*, incorrectly suggesting that the decision rule disparately impacts Group 2 defendants. See main text for calculations.*

dotted line), yielding a true negative rate of $54/78 = 69\%$. Importantly, the true negative rates differ across groups even though the same, risk-conditioned decision rule was applied to each group.

Similarly, the false negative rate is the proportion who are released, among the defendants who *would* violate if released. In Group 1, 19 defendants would violate if released (represented by the $\times$ symbols on the left-hand side of Figure 1). Among these defendants, 7 are released (the $\times$ symbols above the dashed line), resulting in a false negative rate of $7/19 = 37\%$. Moving to Group 2, 22 defendants would violate if released (indicated by the $\times$ symbols on the right-hand side of Figure 1). Among these 22 defendants, 6 are released (those above the dashed line), giving us a false negative rate of $6/22 = 27\%$.

Next, the calculation of Δ requires computing $\bar{\mu}$, which is the expected violation rate that would result from releasing all 200 defendants. Among the 200 defendants depicted in Figure 1, there are 41 defendants who would violate if released (represented by the $\times$ symbols), yielding an overall violation rate of $41/200 = 21\%$. Finally, we compute Δ using the results above:

$$\begin{aligned}\Delta &= (\delta_1^T - \delta_2^T)(1 - \bar{\mu}) + (\delta_1^F - \delta_2^F)\bar{\mu} \\ &= (0.78 - 0.69)(1 - 0.21) + (0.37 - 0.27)(0.21) = 0.1.\end{aligned}$$

The resulting value of $\Delta = 0.1$ suggests disparate impact to the disadvantage of defendants in Group 2. The only way for the judge to reduce Δ is to unfairly detain low risk defendants and release high risk defendants.

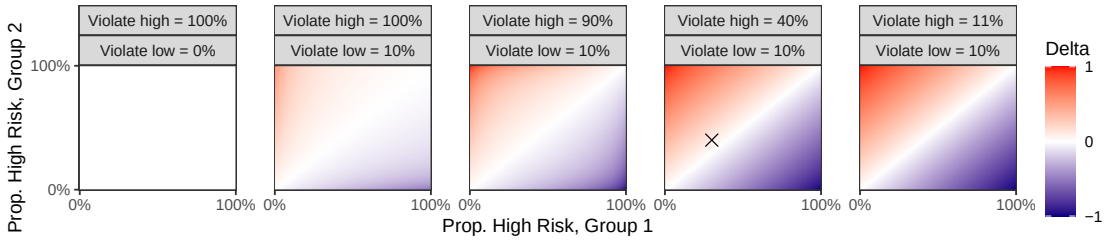

Figure 2: *Extension of the scenario in Figure 1 to different discrete distributions of risk and different violation probabilities. The* $\times$ *symbol panel denotes the scenario from Figure 1. When risk distributions are identical, the* Δ *(Delta) measure of disparity correctly indicates no disparate impact, as indicated by the white diagonal in each panel. The leftmost panel shows that the* Δ *measure correctly indicates no disparate impact when low risk defendants never violate and high risk defendants always violate. However, as the violation probabilities of low- and high-risk defendants move away from the extremes, differences in the underlying distributions of risk result in non-zero values of* Δ*, incorrectly indicating evidence of disparate impact.*

Figure 2 extends the example in Figure 1 to a range of similar scenarios in which a judge only detains high risk defendants. The scenario in Figure 1, in which 30% of Group 1 defendants and 40% of Group 2 defendants are high risk, is denoted by the $\times$ symbol in the fourth panel of Figure 2. Appendix Figure A1 extends this example to continuous distributions of risk.

Overall, these results illustrate the problem of infra-marginality that plagues error rates: Error rates change as a function of the risk distributions. Hence, the Δ measure of disparate impact correctly indicates the absence of disparate impact in only two scenarios: (i) when risk is perfectly predictive, such that high risk defendants *always* carry contraband and low-risk defendants *never* do; (ii) or when the risk distributions between the groups are identical. In all other cases, the Δ measure will indicate disparate impact on Group 2 if Group 2 defendants are, on average, riskier than Group 1 defendants, and vice versa. As the violation probabilities for low risk and high risk defendants move from the extremes, the Δ measure of disparate impact is more sensitive to differences in the distribution of risk across groups. Because the measure does not allow us to draw accurate inferences about the decision rule, we believe it is ill-suited to capture disparate impact. Indeed, it is our view that this and other estimands based on error rates are not appropriate to accurately capture notions of fairness, calling into question their utility [Chohlas-Wood et al., 2023, Simoiu et al., 2017].

3.2 Risk-adjusted disparities

An alternative approach to the measure of disparate impact is risk-adjusted regression. The approach consists of two steps. First, the analyst uses all available data to create a risk-model of the form

$$\texttt{risk}_i = g(X_i), \tag{3}$$

where $\texttt{risk}_i$ is the estimated risk of subject i, g is an arbitrary function, and $\vec{X}$ is a vector of included controls. In the second step, the analyst fits a model of the following (or similar) form:

$$\Pr(\texttt{Y}_i = 1) = \alpha_{\texttt{race}[i]} + \gamma \cdot \texttt{risk}_i, \tag{4}$$

where $\texttt{Y}_i$ is the binary outcome of the decision, $\alpha_{\texttt{race}[i]}$ is an intercept term shared by subjects with the same race or ethnicity as subject i, $\texttt{risk}_i$ is the estimated risk of subject i, as estimated from the first model, and γ is its associated coefficient.

Consider how this approach connects to the legal definition of disparate impact. Assuming that g is sufficiently flexible, the first model reflects the analyst's best attempt to capture an individual's probability that the relevant outcome (e.g., weapon recovery, recidivism or satisfactory job performance) will occur. If the actual decisions made were fully explainable by the individual's risk, the coefficient on $\alpha_{\texttt{race}}$ would be (close to) 0. But if instead the coefficient is significantly different from 0, this suggests that the actual decision rule imposes disparities that are not justified by risk. In this sense, a risk-adjusted regression speaks to the first two elements of a disparate impact claim. It can suggest the existence of an unjustified, adverse impact. At the same time, a risk-adjusted regression itself does not specify a specific, implementable policy, because it does not propose any particular decision rule. As such, it does not fulfill the third element, the showing of an alternative policy with less of an adverse impact.

3.3 Optimized decision making

In a scenario where risk or qualification can be estimated for every individual, the utility-maximizing decision rule is one where a unilateral threshold dictates decision making [cf. Corbett-Davies et al., 2023]. In other words, individuals with estimated risk or qualification above the threshold are selected, and individuals below are not. Among others, this approach has been used by Elzayn et al. [2023] to measure adverse impact under hypothetical risk thresholds. Similar to risk-adjusted regression, the first step consists of estimating a risk model of the form

$$\texttt{risk}_i = g(X_i). \tag{5}$$

After risk has been estimated, individuals are sorted based on their estimated risk. A threshold is drawn such that everyone above the threshold receives the costs/benefits and anyone below the threshold does not. How exactly the threshold is drawn is a matter of policy, and typically reflects some type of constraint. For instance, in defining a reference policy for the auditing practices of the IRS, Elzayn et al. [2023] pick the threshold such that the number of people audited are the same as under current IRS practices. Other possibilities are to draw a threshold such that the risk to public safety or the amount of loans given out are the same as under a current policy. After defining the threshold, the disparities are assessed by comparing the group-specific probability of receiving the cost/benefit.

Consider how this statistical approach relates to disparate impact law. Disparate impact law requires a showing of a feasible, alternative policy that has fewer disparities while

achieving the stated policy goal at least as effectively as the current policy. If such a policy exists, it implies that the (greater) disparities under the current decision rule are avoidable. In this way, disparate impact law can be understood as a search over the policy space for policies that fulfill the before-mentioned criteria. This approach is equivalent to assessing a subset of the policy space for whether it provides less disparate alternatives. Importantly, threshold rules are not a random subset of decision rules. Instead, as Corbett-Davies et al. [2023] and others have shown, threshold decision rules are uniquely optimal among all policies, given estimated risk.

Both risk-adjusted regression and the search for risk-based alternative policies require an estimation of risk, reflected in g. Because the estimation of risk is purely predictive, g can, in principle, be arbitrarily flexible. For instance, risk can be estimated via a random forest or neural network. Similarly, X can contain an arbitrarily large set of covariates.[10] However, as detailed above, disparate impact law requires the plaintiff to propose alternative policies that are feasible and implementable. Depending on the context, it may be argued that such feasibility requires the imposition of constraints, both on g and on X. Take, for instance, frisk decisions. The decision to stop and frisk is often a split-second decision that is made in the moment. If g takes a complex functional form such as a neural network, the model will uncover statistical associations that a police officer who is patrolling the beat might not be able to uncover themselves. Thus, the only way for the officer to meet the standard implicitly set by the use of g would be for them to use the risk model themself, e.g., by feeding a feature vector for the potential suspect into the model and obtaining the prediction. This is not always realistic, and so we may want to confine g to resemble decision making rules that the officer can quickly employ while on patrol. Such concerns are of less relevance, however, if well-resourced actors are making decisions without imminent time constraints. Indeed, some entities are already using complex algorithms, as is the case when the IRS makes its auditing decisions. Allowing g to be flexible is in such contexts is merely akin to a requirement that they use the best available algorithm, which can often be achieved with relative ease. In the next section, we discuss the implementability of threshold rules in more detail.

4 Measuring disparate impact in policing

To illustrate the discussed approaches, we next estimate disparate impact in frisk decisions made by New York City Police Department (NYPD) officers, akin to Jung et al. [2019]. In doing so, we highlight that the liability of police departments under existing disparate impact laws is, as a legal matter, highly theoretical and contested [Tiwari, 2019]. However, legal irrelevance does not imply policy or normative irrelevance.

Until 2013, when the practice was substantially curbed, NYPD officers could stop and question pedestrians given reasonable suspicion of criminal activity. Officers are also permitted to conduct a frisk, or a brief pat down, of a stopped individual's outer clothing if they suspect the individual is armed and dangerous. The practice was often criticized as

[10]The main reason not to include race itself in the risk estimation is that this may constitute disparate treatment. Similarly, in the presence of label bias, addition of certain covariates may reduce the performance of the risk model on the true label [Zanger-Tishler et al., 2023]

racially discriminatory, since stopped individuals were disproportionately likely to be from minority groups.

We begin our analysis with a dataset of all 2.2 million stops recorded by the New York City Police Department between January 1, 2008 and December 31, 2011. For simplicity, we restrict our analysis to the approximately 1.5 million stops of Black and white individuals. In this restricted set, 85% of stops were of Black individuals. Criminal possession of a weapon was the most common reason for conducting a stop, with 28% of stops having this classification. (Table A1 in the appendix provides additional summary statistics for the dataset.) We find that 57% of stopped Black individuals were frisked, compared to 44% of stopped white individuals. This difference in frisk rates is the *prima facie* adverse impact component of a disparate impact claim.[11] To determine whether the observed adverse impact in frisk rates is justified, we first measure risk-adjusted disparities in frisk rates. Then, we construct alternative frisk policies with lower adverse impact and the same or greater efficiency than the status quo policy. In both cases, we find evidence that the NYPD frisk practices imposed a disparate impact on Black individuals.

To generate risk estimates required for both risk-adjusted regression and risk-thresholded decision rules, we fit a model estimating the likelihood that a frisk of a stopped individual will recover a weapon. After subsetting to the individuals in the dataset who were frisked in 2008, 2009, or 2010, we fit a random forest model predicting weapon recovery based on all observed factors that an officer could reasonably account for in their decision to frisk, irrespective of legality.[12] These covariates include the precinct in which the stop occurred, the suspected crime, the reason(s) for the stop, and the stopped individual's gender and race.[13]

For each individual stopped in 2011, we use the fitted risk model to estimate the probability of carrying a weapon—regardless of whether they were frisked. Appendix Figure A3 shows the distribution of estimated risk for each race group. Of course, weapon recovery is only observed among individuals who were frisked, so it is impossible to verify the accuracy of the risk model among individuals who were not frisked. If there exists an unobserved variable that is correlated with both the frisk decision and the likelihood of carrying a weapon, then our risk estimates will suffer from omitted variable bias [Angrist and Pischke, 2008]. For the purposes of illustration, we proceed under the assumption of no omitted variable bias. In other words, we assume that the decision to frisk is ignorable: conditional on observed covariates, the frisk decision is independent of carrying a weapon.[14]

Figure 3 shows, for all stopped individuals, the observed probability of frisk conditional on the estimated risk of carrying a weapon. Across risk levels, Black individuals are substantially more likely to be frisked than white individuals, with a larger gap at higher risk levels. A risk-adjusted regression fit to the data confirms the pattern in Figure 3: Black

[11] Here we focus on disparate impact in frisk decisions, not stop decisions, since we do not have data on those who were not stopped, making the analysis harder for that decision point.

[12] We fit the random forest model in `R` using the `ranger` package. We use 128 trees and the default parameters. One could alternatively fit a more complex risk model, such as a neural network, or a less complex model, such as a regularized logistic regression.

[13] Table A2 in the appendix shows all covariates included in the risk model, along with their relative variable importance. The risk model has an estimated out-of-sample AUC of 0.79.

[14] We note that Jung et al. [2019] outlines a method for assessing the sensitivity of risk-adjusted regression to omitted variable bias in risk estimation.

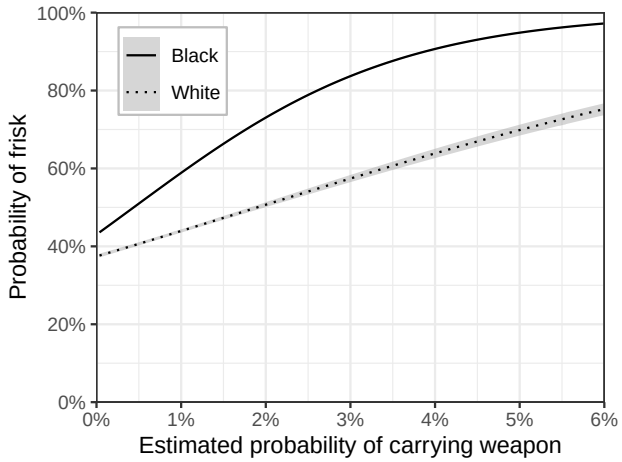

Figure 3: *For stopped individuals, the probability of being frisked as a function of the estimated probability of carrying a weapon, with error bars denoting 95% confidence. Frisk probability is estimated via logistic regression. Black individuals are substantially more likely to be frisked than white individuals with similar estimated risk, with the largest gap at the highest risk levels. Approximately 95% of stopped individuals have an estimated risk of carrying a weapon that is less than 6%.*

individuals had, on average, 1.9 times greater odds of being frisked than white individuals of similar risk. The finding that 57% of stopped Black individuals were frisked compared to 44% of stopped white individuals establishes the *prima facie* adverse impact component of a disparate impact claim. The results of the risk-adjusted regression suggest that this adverse impact is not explained by the estimated risk of recovering a weapon, which is the primary justification for conducting a frisk.

These results suggest that the adverse impact imposed by frisk decisions are not *justified*. In a last step, we turn to the question of whether there exists an implementable alternative frisk policy that has lower adverse impact and is at least as efficient as the status quo frisk policy. Although there is not an agreed-upon definition of efficiency with respect to frisk decisions, for the purpose of this analysis we define a frisk policy as efficient if it recovers at least as many weapons as the status quo policy without increasing the total number of frisks.

To identify an initial efficient threshold policy, we sort all n stopped individuals in descending order by their estimated risk. We then iterate through possible risk thresholds, where the k individuals above each risk threshold t are assumed to be frisked, and the $n-k$ remaining individuals are not. We measure the adverse impact of this policy as the ratio of the resulting frisk rates for Black and white individuals. To sweep out the remaining threshold policies, we iterate over lower values of t until k is approximately the same as the total number of individuals frisked by the status quo policy. At each iteration, we calculate the adverse impact and the expected number of recovered weapons.[15]

Disparate impact law stipulates that the benchmark to which one should measure de-

[15]We calculate this expectation by summing the estimated risk of the k frisked individuals.

cision rates consists of those affected by the decision, or those who could be affected by a change in the way the decision is determined [Carpenter v. Boeing Co., 456 F.3d 1183 (10th Cir. 2006); Hous. Invs., Inc. v. City of Clanton, Ala., 68 F. Supp. 2d 1287 (M.D. Ala. 1999)]. Following these guidelines, we calculate frisk rates using two reasonable benchmark populations. First, we calculate frisk rates by dividing the number of frisks for a given race group by the entire New York City (NYC) population of that race group, as measured by the 2010 U.S. census. The city-level benchmark is intended to be representative of all individuals who could have been frisked by police. Second, we use the total number of stopped individuals in each race group as the denominator of the frisk rate calculation. The second benchmark follows from a narrower perspective of disparate impact in frisk decisions where the affected group consists of just those who were stopped.[16]

Figure 4 shows the adverse impact resulting from the threshold policies derived from the iterative process outlined above. The left panel benchmarks frisk rates to the entire NYC population, while the second benchmarks to just stopped individuals. The dotted line in each panel represents a policy where frisk decisions are determined by whether estimated risk crosses a given threshold. The solid point at the far left of the line indicates the adverse impact observed for the threshold policy that recovers, in expectation, the same number of weapons as the status quo policy.[17] As we sweep across smaller thresholds, the total number of allowed frisks increases. The x-axis shows, for each policy, the number of frisks conducted (k) divided by the total number of frisks observed in the real data (n).[18] Finally, for comparison, the solid arrow on the right side of each panel indicates the adverse impact observed under the status quo frisk policy. Across all thresholds, these threshold policies have lower adverse impact than the status quo. Further, all of these policies are able to recover at least as many weapons as the status quo with fewer frisks. These results show the existence of an equally efficient policy with lower adverse impact, arguably meeting the plaintiff's burden under the third step of a disparate impact claim.

One might, however, argue that such complicated decision rules—which involve complex risk estimation—are not practically implementable. To address this concern, we follow [Goel et al., 2016] and construct more readily implementable decision rules (See "Constructing the simple rule" in the Appendix for the rule construction process). These simple rules account only for the precinct in which the stop occurred, whether the stop occurred in a public housing or transit setting, whether the suspected crime is criminal possession of a weapon, whether the stop was conducted due to the officer observing either a suspicious object or bulge, and whether there were additional sights or sounds indicative of criminal activity. To use these simple rules, officers would only need to add up four small integers, and compare the result to a threshold unique to each combination of precinct and location type. The solid line in Figure 4 shows the adverse impact resulting from threshold policies based on these simple rules. The simple rule threshold policy that conducts the same number of frisks as the status quo, indicated by the rightmost point on the solid line, exhibits substantially

[16]Ultimately, the scope of a hypothetical disparate impact claim would inform the appropriate choice of reference group.

[17]This is the policy with the highest threshold that is arguably as efficient as the status quo policy, so we do not show policies with higher thresholds (i.e., fewer frisks). Analogously, we do not show policies with lower thresholds than the policy that frisks the same number of individuals as the status quo.

[18]Figure A4 in the appendix shows the actual risk threshold for each policy as a function of k/n.

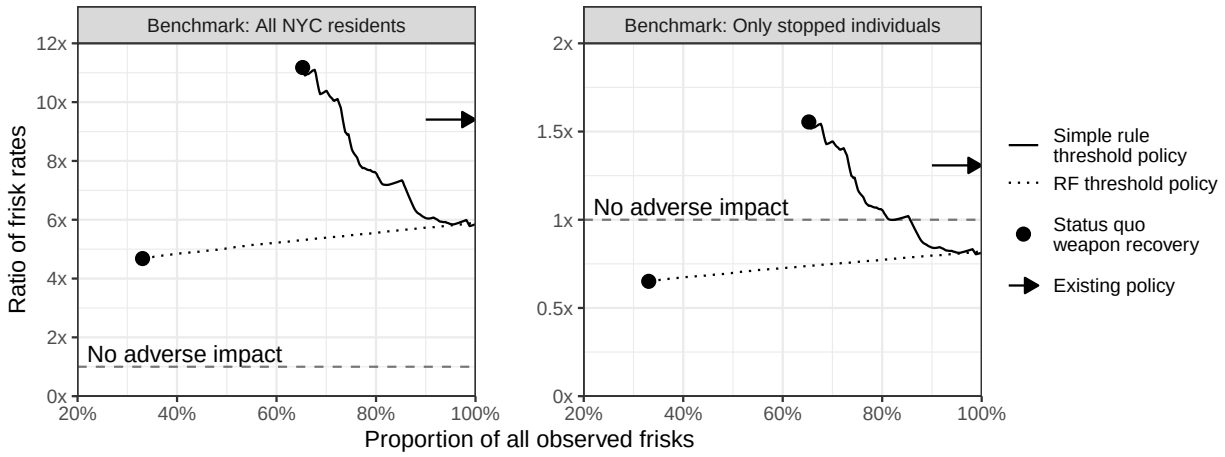

Figure 4: *Estimated adverse impact under hypothetical threshold policies that, compared to the status quo, recover at least as many weapons and result in no more frisks. Adverse impact is measured as the ratio of frisk rates among stopped Black and white individuals. The threshold policy based on risk estimated via a random forest (dotted line) has lower adverse impact across all thresholds than the existing frisk policy (black arrow). However, a more readily implementable threshold policy based on a simple rule (solid line) has lower adverse impact than the status quo only for certain thresholds. At the lowest threshold (solid point), the simple rule threshold policy recovers the same number of weapons as the status quo policy with 40% fewer frisks, but results in substantially higher adverse impact. At the highest threshold, the simple rule policy recovers additional weapons, results in the same number of frisks, and results in substantially lower adverse impact.*

lower adverse impact while also recovering additional weapons. Thus, regardless of whether one uses a complex risk model or a simple, more easily implementable model, it is possible to make frisk decisions with lower adverse impact and greater efficiency, in line with the third requirement of a disparate impact claim.

5 Discussion

In practice, risk-based approaches to disparate impact are only applicable in certain settings. First, there must be a measurable indicator of a successful decision. As an example, consider the pretrial setting. In most jurisdictions, the primary justification for pretrial detention is minimizing the risk of failing to appear or committing new criminal activity. The existence of a pretrial violation is a concrete way to assess whether a release decision is "successful". In other domains, such as college admissions, it is not immediately clear how to denote a successful decision. Second, one must be able to estimate risk accurately. This typically means that decision rates must be high enough such that there exists a sufficient number of individuals from which to estimate risk. Additionally, as accurate risk estimation often rests on the strength of the ignorability assumption, the fitted risk model should incorporate as many of the variables observed by the decision maker as possible. If there are unobservable

variables that are highly predictive of both the decision itself and the success of the decision, the risk model may suffer from severe omitted variable bias. Finally, the proposed risk-based alternative policies must be implementable. In the stop-and-frisk setting, where the decision to frisk is made in a matter of seconds, even a simple rule could be deemed as impossible to realistically implement. For less time-constrained decisions, such as pretrial detention or tax auditing, risk estimates can be generated well in advance of decisions.

In addition, we note that, in this study, we take disparate impact law as given and consider empirical strategies in relation to current legal analysis. But we believe the current law on disparate impact has many shortcomings itself. Among others, disparate impact law's focus on raw disparities can lead decision makers to forego policies that are ultimately favorable to the minority group. For instance, a policy that is strictly beneficial to both the minority and the majority group, but that benefits the majority group more than the minority group, would not need to be enacted under disparate impact law because it *increases* the disparities between the groups.

To illustrate with a numeric example, consider a hypothetical scenario under which the current policy has police officers frisk individuals if they made 'furtive movements.' Under this policy, the officer stops 100 of 10,000 Black citizens a year, and 1,000 of 100,000 white citizens. An analysis shows that, although officers do not act with discriminatory intent, 'furtive movements' is not predictive of weapon recovery. Removing this requirement would thus reduce the number of Black citizens stopped by 50, and the number of white citizens stopped by 600, without meaningfully affecting the weapon recovery rate. In this scenario, the current policy has a frisk rate of 1% for both Black and white citizens. Under the new policy, the frisk rate is reduced to 0.5% for Black citizens and 0.4% for white citizens. But although the new policy decreases the absolute number of both Black and white citizens who are frisked, it *increases* the relative disparity between Black and white citizens from 0.0 to 0.1 percentage points. Under disparate impact law, the new policy need not be implemented, given that it does not decrease the disparities between the two groups.

The example helps clarify the focus of disparate impact law, and how it might differ from other welfare perspectives on fairness. Disparate impact law is primarily concerned with unjustified, differential treatment between the majority and the minority group. However, it is not a mandate to improve the welfare of the minority group, even if that can be done in a costless way. From a welfarist perspective, this might seem problematic, especially in settings where there is no budget constraint.

Additionally, the reference population from which action rates are calculated should, in theory, consist of those who are subjected to the practice in question [Carpenter v. Boeing Co., 456 F.3d 1183 (10th Cir. 2006); Hous. Invs., Inc. v. City of Clanton, Ala., 68 F. Supp. 2d 1287 (M.D. Ala. 1999)]. In practice, though, it is often unclear what the relevant reference population should be. Furthermore, data for certain reference populations may be inaccessible. For example, in the case of lending, one might propose a reference population of all eligible individuals who applied for a loan from the institution in question. However, a more ideal population would be all individuals who *would have* been eligible for a loan, regardless of whether they actually applied. But, the size of this larger group may not be be estimable, in which case the smaller group would be an appropriate reference population so long as it is sufficiently representative of the affected individuals [Frazier v. Consol. Rail Corp., 851 F.2d 1447 (D.C. Cir. 1988)]. Courts have also permitted

reference populations that subsume the affected population, once again so long as the larger population is sufficiently representative [E.E.O.C. v. Joint Apprenticeship Comm. of Joint Indus. Bd. of Elec. Indus., 186 F.3d 110 (2d Cir. 1999)].

6 Conclusion

In this paper, we have discussed statistical approaches for assessing disparate impact. Our analysis suggests that recent estimators centered on error rates capture neither legal nor normative notions of disparate impact. While risk adjusted regressions can help document the existence of unjustified disparities, a concrete, optimal alternative policy can be derived by sorting individuals based on their estimated risk and defining a decision threshold. As we have shown for the example of stop and frisk decisions by the NYPD between 2008 and 2011, this approach relies analysts to formulate alternative, less disparate, implementable policies even in scenarios where decision makers have constrained information or time. We hope that this research will positively contribute towards a current trend to broaden conceptions of discrimination in empirical research.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

David Arnold, Will Dobbie, and Peter Hull. Measuring racial discrimination in algorithms. In *AEA Papers and Proceedings*, volume 111, pages 49–54, 2021.

David Arnold, Will Dobbie, and Peter Hull. Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9):2992–3038, 2022.

Ronen Avraham, Kyle D Logue, and Daniel Schwarcz. Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195, 2013.

Ian Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

Ian Ayres. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of "included variable" bias. *Perspectives in Biology and Medicine*, 48(1): 68–S87, 2005.

Ian Ayres. Testing for discrimination and the problem of "included variable" bias. *Yale Law School*, 2010.

E Jason Baron, Joseph J Doyle Jr, Natalia Emanuel, Peter Hull, and Joseph P Ryan. Racial Discrimination in Child Protection. 2023.

Gary S Becker. *The Economics of Discrimination*. 1957.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

J Aislinn Bohren, Peter Hull, and Alex Imas. Systemic Discrimination: Theory and Measurement. 2022.

George S Bridges and Sara Steen. Racial disparities in official assessments of juvenile offenders: Attributional stereotypes as mediating mechanisms. *American sociological review*, pages 554–570, 1998.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

William Cai, Johann Gaebler, Justin Kaashoek, Lisa Pinals, Samuel Madden, and Sharad Goel. Measuring racial and ethnic disparities in traffic enforcement with large-scale telematics data. *PNAS Nexus*, 1(4), 2022.

Dallas Card and Noah A Smith. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3:34, 2020.

Anupam Chander. The racist algorithm. *Michigan Law Review*, 115:1023, 2016.

Alex Chohlas-Wood, Madison Coots, Sharad Goel, and Julian Nyarko. Designing equitable algorithms. *Nature Computational Science*, 3(7):601–610, 2023.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Sam Corbett-Davies, J Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 2023.

Stephen Demuth. Racial and ethnic differences in pretrial release decisions and outcomes: A comparison of hispanic, black, and white felony arrestees. *Criminology*, 41(3):873–908, 2003.

Stephanie Holmes Didwania. Discretion and disparity in federal detention. *Nw. UL REv.*, 115:1261, 2020.

Edward C Dillon Jr, Juan E Gilbert, Jerlando FL Jackson, and LJ Charleston. The state of african americans in computer science-the need to increase representation. *Computing Research News*, 21(8):2–6, 2015.

Ellen A Donnelly and John M MacDonald. The downstream effects of bail and pretrial detention on racial disparities in incarceration. *J. Crim. l. & CriminoloGY*, 108:775, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Hadi Elzayn, Evelyn Smith, Thomas Hertz, Arun Ramesh, Jacob Goldin, Daniel E Ho, and Robin Fisher. Measuring and mitigating racial disparities in tax audits. 2023.

Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, 9(1):26–48, 2022.

Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 10(1): 365–394, 2016.

Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 169(12):883–884, 2018.

Joshua Grossman, Sabina Tomkins, Lindsay C Page, and Sharad Goel. The Disparate Impacts of College Admissions Policies on Asian American Applicants. 2023.

Ben Grunwald, Julian Nyarko, and John Rappaport. Police agencies on Facebook overreport on Black suspects. *Proceedings of the National Academy of Sciences*, 119(45), 2022.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 2020.

Aziz Huq. Racial equity in algorithmic criminal justice. *Duke Law Journal*, 68, 2019.

Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167–189, 02 2023.

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*, 2019.

Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 64–70, 2020.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 576–586, New York, NY, USA, 2021. Association for Computing Machinery.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

Sandra G Mayson. Bias in, bias out. *The Yale Law Journal*, 128(8):2218–2300, 2019.

Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

Christi Metcalfe and Ted Chiricos. Race, plea, and charge reduction: An assessment of racial disparities in the plea process. *Justice Quarterly*, 35(2):223–253, 2018.

Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine*, 3(1):1–8, 2020.

M Marit Rehavi and Sonja B Starr. Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6):1320–1354, 2014.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

Alisa Tiwari. Disparate-impact liability for policing. *The Yale Law Journal*, pages 252–306, 2019.

John Wooldredge. Distinguishing race effects on pre-trial release and sentencing decisions. *Justice Quarterly*, 29(1):41–75, 2012.

Michael Zanger-Tishler, Julian Nyarko, and Sharad Goel. Risk scores, label bias, and everything but the kitchen sink. *arXiv preprint arXiv:2305.12638*, 2023.

A Appendix

Table of Contents

Figure A1 is an extension of Figure 2 to a setting where risk is continuously distributed. Specifically, risk is parameterized by a beta distribution with a fixed variance and mean between 0 and 1.

Table A1 shows summary statistics for the dataset used in the main analysis.

Table A2 shows the relative variable importance for the random forest risk model fit to individuals frisked in 2008, 2009, or 2010.

Table A3 shows the distribution of estimated risk for individuals stopped in 2011.

"Constructing the simple rule" outlines the process used to construct the simple rule risk model to which the random forest risk model is compared in Figure 4.

Figure A4 shows the numeric risk thresholds for the policies illustrated in Figure 4 as a function of the number of frisks conducted.

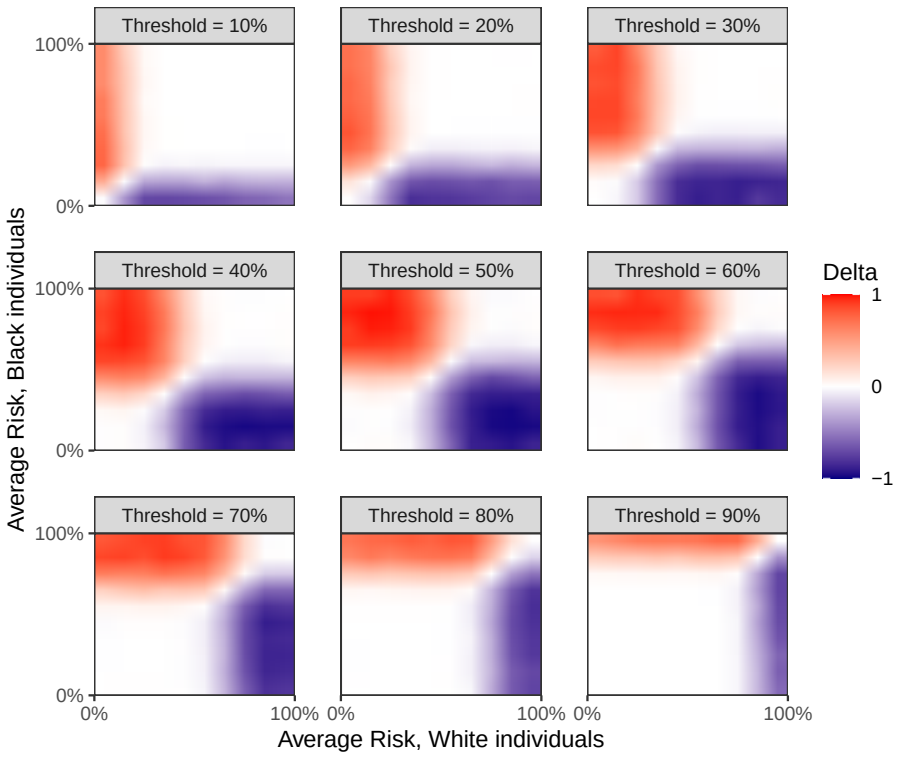

Figure A1: *Extension of the scenario in Figure 1 to continuous risk. In this example, risk follows a beta distribution with a mean between 0 and 1 and a fixed variance of 0.01. Each panel shows values of* Δ *(Delta) calculated under a fixed search threshold for synthetic groups of Black and white stopped individuals, with risk values randomly generated from a beta distribution with the mean specified on the corresponding axis. As in Figure 1, each panel has a white diagonal line, indicating that the* Δ *measure is correctly 0 when risk distributions are identical. When the mean of both risk distributions is far from the search threshold,* Δ *is also close to 0. However, when the mean of either risk distribution is close to the search threshold, the calculated value of* Δ *is more sensitive to small changes in the shape of the risk distribution(s) whose mean is close to the search threshold.*

Variable	All	Black	White
Num. stops	1,462,967	1,237,469	225,498
Prop. all stops	1	0.85	0.15
Num. frisks	808,389	709,068	99,321
Frisked	0.55	0.57	0.44
Weapon if frisked	0.02	0.02	0.04
Female	0.07	0.07	0.1
Under 25	0.54	0.54	0.5
CPW stop	0.28	0.31	0.11
Housing stop	0.18	0.2	0.04
Transit stop	0.08	0.08	0.07
Suspicious obj.	0.02	0.02	0.04
Suspicious bulge	0.09	0.1	0.04
Addl. sights/sounds	0.02	0.02	0.04

Table A1: *Summary statistics for the data used in the main analysis. 85% of stopped individuals in the analysis are Black. Black individuals are, on average, more likely to be frisked: 57% of stopped Black individuals are frisked, compared to 44% of stopped white individuals. However, frisked white individuals are more likely to be carrying a weapon: 4% of frisked white individuals carry a weapon, compared to 2% of frisked Black individuals. The most common reason for a stop is suspicion of criminal possession of a weapon (CPW). 31% of stopped Black individuals are stopped for this reason, compared to 11% of stopped white individuals.*

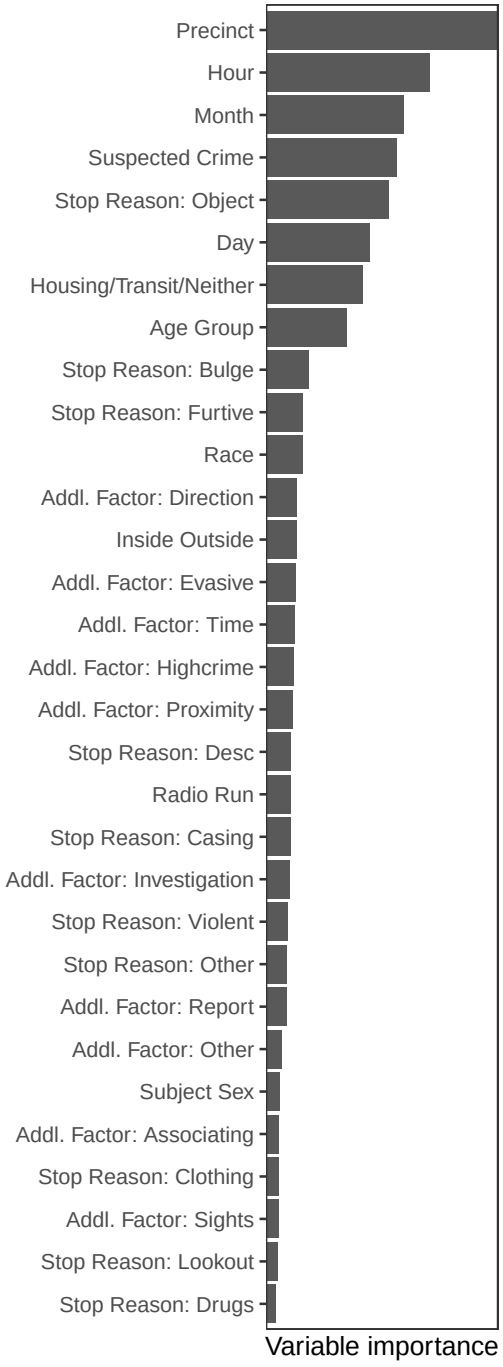

Figure A2: *Relative variable importance for the covariates included in the random forest risk model trained on individuals frisked in 2008, 2009, or 2010. Precinct and time fixed effects have high importance, along with stop location, the suspected crime, and the stop reason of a suspicious object.*

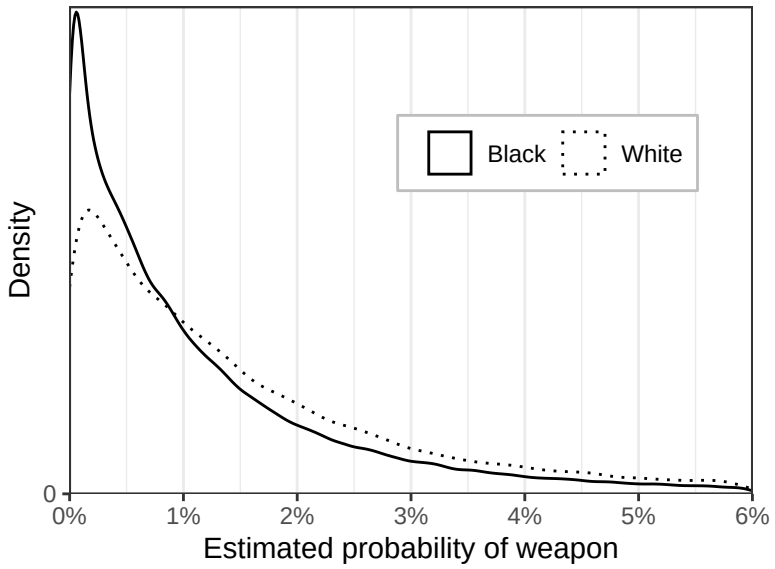

Figure A3: *Distribution of estimated risk of carrying a weapon among individuals stopped in 2011. On average, stopped white individuals are more risky than stopped Black individuals. The majority of stopped individuals have an estimated likelihood of carrying a weapon that is less than 2%.*

Constructing the simple rule

To construct the simple rule risk model, we begin with the same set of covariates as those listed in Table A2. Goel et al. [2016] found that the covariates most predictive of recovering a weapon during a frisk were a suspicious object, a suspicious bulge, and additional sights and sounds of criminal activity. Goel et al. [2016] restricted their analysis to just the stops where the suspected crime was criminal possession of a weapon. We consider all stops, and so we include an additional covariate that indicates whether the suspected crime was criminal possession of a weapon. For simplicity, we use just these four covariates in our simple rule, though one could more methodically choose covariates using the procedure outlined in Jung et al. [2020].

Risk varies substantially with precinct and location type. As such, we fit a logistic regression model to individuals frisked in 2008, 2009, or 2010 that estimates the likelihood of recovering a weapon using the precinct, location type, and the four covariates noted above. We fit this model only on frisked individuals, as the weapon recovery outcome is unknown for individuals who are not frisked. With this fitted model in hand, we multiply the fitted coefficients of the four chosen covariates by 10 to put the coefficients on an approximate integer scale, and then round the four coefficients to the nearest integer. This procedure gives us rounded coefficients of 18 for suspected criminal possession of a weapon, 24 for a suspicious object, 8 for a suspicious bulge, and 5 for additional sights and sounds of criminal activity. Using these four rounded coefficients, we calculate a score from 0 to 55 for each stopped individual. Finally, we fit an additional logistic regression model that predicts weapon recovery using just this score and the precinct. This model is fit just to frisked individuals. This final fit provides the optimal coefficients for each precinct and location type.

To operationalize the simple rule, the New York City Police Department could select a risk threshold above which officers are permitted to frisk. Alternatively, each precinct could select its own risk threshold. Using the final fitted model, NYPD would determine, for each combination of precinct and location type, the minimum score needed for the estimated risk of weapon recovery to exceed the desired threshold. Officers assigned to each combination of precinct and location type would be notified of the relevant threshold score. When stopping an individual, officers could quickly calculate the individual's score using the simple rule, and then choose to frisk if the score exceeds the known threshold.

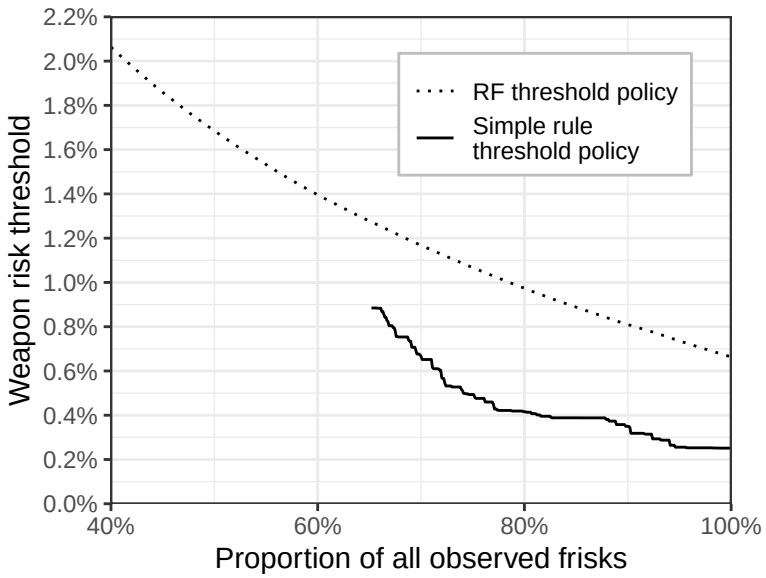

Figure A4: *Risk thresholds for the policies illustrated in Figure 4. The lines begin at the risk threshold for which the policy recovers the same number of weapons as the status quo, in expectation. The simple rule sorts individuals by risk less accurately than the random forest model, so requires lower risk thresholds (i.e., more frisks) in order to recover the same number of weapons in expectation.*