Racial Profiling? Detecting Bias Using Statistical Evidence*

Nicola Persico

New York University

November 4, 2008

Abstract

We review the economics literature which deals with identifying bias, or taste for discrimination, using statistical evidence. A unified model is developed which encompasses several different strategies which have been studied in the literature. We also discuss certain more theoretical questions concerning the proper objective of discrimination law.

KEYWORDS: RACIAL PROFILING, DISCRIMINATION, BIAS.

---

*Department of Economics, 19 W. 4th Street, New York, NY 10012. Email: nicola@nicolapersico.com. Support from NSF Award ID 0617507 is gratefully acknowledged. Thanks to Hanming Fang, Claudia Goldin, Jeff Grogger, Dragan Ilic, Cecilia Rouse, and especially Charles Manski for improving the manuscript.

1 Introduction

The term "racial profiling" typically refers to discretionary enforcement practices which have a disparate impact by race. However, the term is also used more broadly, in reference to a number of situations in which one or more treators chooses whom to treat among a number of agents with heterogeneous characteristics. This broader acception of the term "profiling" is sometimes applied to disparities in medical care, in lending, in jury selection, etc., as well as disparities by gender and other characteristics. The term racial profiling is derogatory—it is usually taken to denote a reprehensible, possibly illegal practice. Yet among the public there is no consensus about how broadly it should be applied, because people have different views about what exactly should be (or is) illegal. At one extreme, some people tend to think that most disparities in outcomes should be deemed illegal. The other extreme is the view that only disparities reflecting an invidious bias, or intentional discrimination, should be illegal.

The law is fairly clear—at least in its broad principles. Discrimination law in the United States generally does not prohibit disparities in outcomes (referred to, in legal parlance, as disparate impact), as long as they do not reflect an intent to discriminate. The expression "intentional discrimination" has a specific legal meaning: it is taken to mean that the treator engaged in disparate treatment "because of," not merely "in spite of," its adverse effects upon an identifiable group. A mere awareness of the consequences of an otherwise neutral policy does not suffice.[1]

[1] Citing *Personnel Administrator of Massachusetts v. Feeney*, 442 U.S. 256, 279 (1979). There are two nuances to this statement. First, there is the question of allocating the burden of proof. In some areas of the law (employment law is one of them), once the plaintiff shows disparate impact, it is up to the defendant

The practical difficulty with translating this broad legal principle into practice is that it is often not obvious how intentional discrimination could be distinguished from other (legitimate) sources of disparity. Typically, when attempting to infer discrimination from statistical data, a key difficulty arises from "missing variables:" among the legitimate sources of disparate impact are productive characteristics of the treated, which are potentially correlated with race, but may not be observed by the researcher. Because statistical data are unlikely to record all productive characteristics of the treated, this potential problem is pervasive. Thus many view the task of proving intentional discrimination using statistical data as exceedingly difficult, perhaps impossible absent evidence of explicitly discriminatory statements (typically verbal) by the defendant.

A segment of the economics literature is concerned precisely with identifying intentional discrimination from statistical data in the presence of confounding factors such as missing variables.[2] In this paper we provide an overview of a number of papers which deal with this subject. In order to give a coherent account of this part of the literature, we develop a unified model which encompasses many instances of racial profiling which have been studied in the literature. Writing down the general model will force us to be precise about what features are common to this literature, and what differences require the use of various identification strategies.[3] As the model is specialized in Sections 4 through 7, we briefly discuss some key papers in the literature. In Section 8 we step back from the question of identification and

---

to show that the disparities do not reflect intentional discrimination. Second, the law permits regulations issued by government agencies to forbid disparate impact.

[2]Pair-audit studies are designed precisely to side-step this issue. Their strategy is to design "experiments" where the researcher eliminates unobserved characteristics by using tester pairs that are identical in all productive characteristics and differ only by race. We will not review the literature on pair-audit studies; Heckman and Siegelman (1993) provide a critical assessment of this methodology.

[3]Ayres (2002) and Todd (2006) also discuss identification of bias in racial-profiling environments.

turn to more theoretical questions concerning the proper objective of discrimination law.

2 A model

In this section we sketch a model that underlies several of the identification results in the literature. Let's start with the actors. There are two types of actors in the model; the potential discriminators, and those potentially discriminated against. We shall refer to the first group as the *treators*, and to the second group as the *treated*.

The treated are modeled as a mass of atomistic agents, distinguishable in the eyes of a treator by their characteristics $g \in \mathcal{G}$. For example, if the treator is a lender who can observe the race, education, and credit score of the applicant, then each g represents a vector of the applicant's race, education and credit score. Race is a salient characteristic in our analysis, and for expositional ease we shall assume that there are only two races, African-American and white. The set of possible characteristics observed by the treators is denoted by $\mathcal{G}$.

A treator chooses an action $a_g \geq 0$ for each group, subject to the vector $\{a_g\} \in \mathcal{A}$. The actions represent the extent of treatment that is applied to group g. The set $\mathcal{A}$ represents the set of possible actions the treator can undertake. In the case of lending, for instance, a_g could represent how many members of group g receive a loan, and the set $\mathcal{A}$ would capture the constraint that the total amount lent to all groups cannot exceeds the lender's budget.

The treator "legitimately" cares about achieving an objective which is represented by the function $\pi(a_g, \mathbf{a}_g, g)$.[4] While the legitimate objective may depend on race, it only does so

[4] This raises the question of what is a "legitimate" objective. Often this is clear in a specific context.

because of legitimate (i.e., productive) reasons. Note that the objective also depends on the scalar $\mathbf{a}_g$. This scalar captures the aggregate behavior of other treators beyond the one we are studying. In the case of the lender, the legitimate objective could be maximizing the profits from lending to group g, and $\mathbf{a}_g$ could represent the amount of credit extended by the credit sector as a whole to applicants in group g. We shall assume henceforth that $\pi(\mathbf{a}_g, g)$ is decreasing in $\mathbf{a}_g$.[5]

For notational ease, we shall proceed under the assumption that g's and a_g's are continuous real variables, that the G's are measurable, and that $\pi(a_g, \mathbf{a}_g, g)$ is a continuous function of its arguments.[6]

The treator's objective function potentially comingles the "legitimate" part $\pi(a_g, \mathbf{a}_g, g)$ with illegitimate bias. We model this bias as a multiplicative group-specific coefficient $\beta(g)$, and assume that the treator chooses the vector $\{a_g\}$ in order to optimize the function[7]

$$\int \beta(g)\, \pi(a_g, \mathbf{a}_g, g)\, dg. \tag{1}$$

The coefficient $\beta(g) \geq 0$ captures a psychic cost, of key interest but unobserved, which leads

A question arises when the treator's objective is not race-neutral not because of any treator's taste for discrimination, but because of a taste for discrimination by third parties. For example, a store owner hiring clerks may decline to hire minorities owing to the customers' discriminatory taste. In this case, the law holds that the store owner should ignore the customer's discriminatory tastes. Another question arises when the legitimate objective, even though race-neutral, leads to inefficient disparities by race. We will return to this point in Section 8.

[5] Note that we do not explicitly endow the treated with actions. The possible response of the treated to the treatment is embodied in the function $\pi(a_g, \mathbf{a}_g, g)$.

[6] These assumptions allow us to write integrals, but nothing in the analysis that follows rests on these assumptions. In particular, the analysis would carry through if the g's were elements of a finite set with no cardinal or ordinal structure.

[7] Throughout the paper we use the convention that, when the domain of integration is not specified, it is intended to be the full set of characteristics $\mathcal{G}$.

the treator to privilege the interest of certain subgroups of the population. In the case of lending, for example, a low $\beta(g)$ means that the lender discounts the flow of profits coming from group g, which makes the lender less inclined to lend to that group. In the literature the parameter $\beta(g)$ is often referred to as "taste for discrimination."[8] We use $\beta(g)$ to capture the legal concept of intentional discrimination.

We do not assume that we are able to observe the full set of characteristics g observed by the treator. Rather, we assume that we are able to observe whether a treated agent's characteristics belong to a partition (with generic element G) of the set $\mathcal{G}$ of all possible characteristics. Suppose, for instance, that credit applicants differ along just two dimensions: their race A or W, and the inherent value H or L of their business idea. In this case the full set of characteristics observed by the lender is $\mathcal{G} = \{(A,H),(A,L),(W,H),(W,L)\}$. Suppose we are only able to tell the race of successful credit applicants, but not the value of their business idea. The partition we observe is then composed of two elements $G_1 = \{(A,H),(A,L)\}$ and $G_2 = \{(W,H),(W,L)\}$. For each successful applicant we are able to tell whether s/he belongs to G_1 or G_2, but not his/her full set of characteristics. Thus, in the language of econometric theory, we may have missing variables. This happens when the partition we observe is very coarse. We shall denote by R the set that includes all groups with race equal to $R \in \{A,W\}$, so that in the previous example we would write $G_1 = A$ and $G_2 = W$. We will assume that, at a minimum, we are able to distinguish the race of the

[8] See Becker (1973). An alternative way of introducing a taste for discrimination is through an additive parameter. In this alternative formulation, the agent maximizes something like $\int [\pi(\mathbf{a}_g, g) - \beta(g)]\, a_g \, dg$. This is the formulation adopted by Knowles *et al.* (2001) and Anwar and Fang (2006), for example. The analysis we present would go through with minor adaptations in the additive setup. The main advantage of the multiplicative model we use is that it makes it easy to subsume the set of environments studied in Section 7.

treated in our sample. Formally, then our partition is always at least as fine as the partition $\{A, W\}$.

The observability restrictions we take on are compactly summarized as follows. We assume that, for all elements G of a given partition of $\mathcal{G}$, we are able to observe

$$\begin{aligned} &\int_G \pi\left(a_g^*, \mathbf{a}_g, g\right) a_g^* \, dg, \quad \text{and} \\ &\int_G a_g^* \, dg, \end{aligned} \tag{2}$$

where a_g^* denotes the optimal (for the treator) choice of actions. In the lending example, the first quantity represents total profitability from loans to group G, the second quantity represents total money lent to group G. In the following sections we shall see how far we can go by observing only (2), and how other observables (typically, sources of exogenous variation) can aid in identifying bias.

As mentioned above, the fact that we do not observe each g separately captures the fact that we have missing variables. We will not need to assume that the missing variables are independent of the variables we do observe (race, in the example). In a similar vein, we make no use of any cardinal or order structure on the g's or even the G's for estimation. Thus, for example, we have no information (and thus use none) about the g's or the G's for which $a_g^* = 0$. In this sense, the identification strategies we will cover work differently from the conventional econometric selection models.[9]

An implication of our observability restriction is that we may have no information whatever

[9] See, e.g., Amemiya (1985).

about treated agents for which $a_g = 0$. In the case of lending, for example, this means that any information we have only reflects applicants who receive some (positive) amount of credit. This assumption may be too restrictive; in some cases, records are kept about those who are not treated (lending may actually be one of those cases). In other circumstances, however, this assumption is appropriate. In enforcement discrimination, for example, information about those who are not subjected to enforcement is often not available.[10]

Our goal is to identify the function $\beta(g)$.[11] Naturally, our ability to do so will partly depend on how coarse our partition is vis-a-vis the variability in the function $\beta(\cdot)$. Our identification task is easier if we place some restriction on $\beta(g)$. The literature typically proceeds under the assumption that

$$\beta(g) = \beta(R) \text{ for all } g \subset R \tag{3}$$

that is, that bias is constant across all characteristics (observable and not) except for race. We too shall make this assumption.

An implicit assumption is that *all* our observations come from solving the optimization problem studied in this section. In some cases, some of the observations may actually be generated by a different process, a fact which poses additional challenges for identification.[12]

[10] For instance, data concerning the New York Police Department's practice to "stop and frisk" pedestrians do not include information on pedestrians who were not stopped (see Gelman *et al.* 2007).

[11] Up to a linear transformation, of course.

[12] Hernandez-Murillo and Knowles (2004) develop statistical methods to deal with a vehicular search dataset in which an unknown fraction of the observations are generated by a non-discretionary search process (for example, the search was executed incident to an arrest and thus prescribed by police regulations).

3 Examples

The model we discussed fits, at a broad level, a number of important applications. Below, we list some of them.

Lending discrimination. The treator is a single lender, the treated are the population of credit applicants to that lender. a_g (resp., $\mathbf{a}_g$) represents the amount of credit extended by that lender (resp., by the entire credit sector) to applicants with characteristics g. The constraint $\{a_g\} \in \mathcal{A}$ captures the lender's budget constraint—the total amount of funds available to the lender is fixed. The function $\pi(a_g, \mathbf{a}_g, g)$ captures the expected profit when a loan of size a_g is extended to a member of group g. Imposing the special structure $\pi(a_g, \mathbf{a}_g, g) = \pi(\mathbf{a}_g, g)\, a_g$ captures competitive behavior on the lender's part. The assumption that $\pi(\mathbf{a}_g, g)$ is decreasing in $\mathbf{a}_g$ means that the expected profit on the marginal dollar lent decreases as the sectors directs more credit opportunities to member g applicants. (This could be, for example, because interest rates charged in equilibrium decrease).

Hiring discrimination. The treator is a school principal, the treated are applicants for a given number of teaching positions. a_g represents the number of applicants with characteristics g who are hired. The constraint $\{a_g\} \in \mathcal{A}$ captures the fact that only as many teachers can be hired as the number of open positions. Suppose that wages are fixed exogenously to be the same for all g (for example by a rigid sectoral contract), and an unbiased principal maximizes the probability of hiring an effective teacher. Then $\pi(a_g, \mathbf{a}_g, g) = \pi(g)\, a_g$, where $\pi(g)$ captures the probability that an applicant of group

g is an effective teacher.

Health care discrimination. The treator is a primary physician, the treated are her patients. a_g represents the number of patients with a collection of symptoms$\times$characteristics g who are referred to a specialist for further evaluation. Suppose an unbiased physician refers for further testing those patients whose symptoms indicate a sufficiently high probability of having a disease. Then $\pi(a_g, \mathbf{a}_g, g) = \pi(g) a_g$, where $\pi(g)$ captures the probability that a patient of group g has the disease.

Enforcement discrimination. The treator is a single police officer, the treated are the citizens under his jurisdiction. a_g (resp., $\mathbf{a}_g$) represents the number of citizens with characteristics g who are searched by that officer (resp., by the entire police force). The constraint $\{a_g\} \in \mathcal{A}$ captures the fact that a police officer can only search so many citizens in a day. Suppose the police officer maximizes the probability of finding contraband. If a single police officer has negligible aggregate impact, we may assume that $\pi(a_g, \mathbf{a}_g, g) = \pi(\mathbf{a}_g, g) a_g$, where $\pi(\mathbf{a}_g, g)$ represents the probability of finding contraband when searching a member of group g. The dependence on $\mathbf{a}_g$ captures the idea that, if the entire police force focusses on group g, then members of that group become less likely to carry contraband. We will also consider the specification in which the treator is the entire police force, and so $\pi(a_g, \mathbf{a}_g, g) = \pi(a_g, g)$. In this case, $\pi(a_g, g)$ could capture the aggregate crime committed by members of group g.

Selective prosecution. The treator is a prosecutor, the treated are the cases under his jurisdiction. a_g represents the number of accused with characteristics g who are prosecuted by that officer. The constraint $\{a_g\} \in \mathcal{A}$ captures the fact that the prosecutor can only prosecute a given number of cases. Suppose an unbiased prosecutor maximizes the probability of conviction. Then $\pi(a_g, \mathbf{a}_g, g) = \pi(g)\, a_g$, where $\pi(g)$ captures the probability of conviction when prosecuting a member of group g.

Sentencing discrimination. The treator is a judge, the treated are defendants who are before him. a_g represents the fraction of defendants of group g who are convicted by that judge. Suppose an unbiased judge maximizes the probability of convicting the guilty. Then $\pi(a_g, \mathbf{a}_g, g) = \pi(g)\, a_g$, where $\pi(g)$ captures the probability that a defendant of group g is guilty.

4 Identification without variation

4.1 Basic model: partial equilibrium

In the basic version of the model we focus attention on only one treator. In addition, we impose two restrictions on our general model. First, we assume that the objective function is linear in the treator's action:

$$\pi(a_g, \mathbf{a}_g, g) = \pi(\mathbf{a}_g, g) \cdot a_g.$$

The fact that a_g enters the objective function multiplicatively embodies the idea that the treator's behavior has a negligible impact on the environment. We take $\mathbf{a}_g$ as an exogenous parameter; this assumption will be relaxed in the next section.

Second, we assume that the treator's action set is given by

$$\mathcal{A} = \left\{ \{a_g\} \text{ s.t. } \int a_g \, dg \leq C \right\}. \tag{4}$$

This constraint can be seen as an aggregate resource constraint, where a_g represents the amount of resources devoted to group g and C represents the total amount of resources available to the treator. A key feature of this constraint is that there is perfect substitutability between efforts devoted to different groups. Both assumptions will be relaxed, at some cost, in later sections.

Given our assumptions, the treator's optimization problem is

$$\max_{\{a_g\}} \int \beta(g) \, \pi(\mathbf{a}_g, g) \, a_g \, dg \quad \text{s.t.} \quad \int a_g \, dg \leq C \tag{5}$$

$$a_g \geq 0 \text{ for all } g.$$

The solution $\left\{a_g^*\right\}$ to this constrained maximization problem maximizes the Lagrangean

$$\begin{aligned} \mathcal{L}(\{a_g\}, \lambda_0) &= \int \beta(g) \, \pi(\mathbf{a}_g, g) \, a_g \, dg - \lambda_0 \left[\int a_g \, dg - C \right] \\ &= \int \left[\beta(g) \, \pi(\mathbf{a}_g, g) - \lambda_0 \right] a_g \, dg + \lambda_0 C, \end{aligned}$$

subject to the constraint that each $a_g \geq 0$. If an optimal a_g^* is strictly positive and finite, then maximization of the Lagrangean implies that

$$\left[\beta\left(g\right)\pi\left(\mathbf{a}_g,g\right)-\lambda_0\right]=0. \tag{6}$$

If the optimal a_g^* is zero then $\left[\beta\left(g\right)\pi\left(\mathbf{a}_g,g\right)-\lambda_0\right]\leq 0$. In either case we may write $\pi\left(\mathbf{a}_g,g\right)a_g^*=\lambda_0\left(1/\beta\left(g\right)\right)a_g^*$. Integrating over any G yields

$$\int_G \pi\left(\mathbf{a}_g,g\right)a_g^*\;dg=\;\lambda_0\int_G \frac{1}{\beta\left(g\right)}a_g^*\;dg.$$

Evaluate at G and at G' and form the ratio to get[13]

$$\frac{\int_G \pi\left(\mathbf{a}_g,g\right)a_g^*\;dg}{\int_{G'} \pi\left(\mathbf{a}_g,g\right)a_g^*\;dg}=\frac{\int_G \frac{1}{\beta(g)}a_g^*\;dg}{\int_{G'} \frac{1}{\beta(g)}a_g^*\;dg.} \tag{7}$$

For all $G\subset W$ and $G'\subset A$ we have, in light of assumption (3),

$$\frac{\beta\left(A\right)}{\beta\left(W\right)}=\frac{\frac{\int_G \pi(\mathbf{a}_g,g)a_g^*\;dg}{\int_G a_g^*\;dg}}{\frac{\int_{G'} \pi(\mathbf{a}_g,g)a_g^*\;dg}{\int_{G'} a_g^*\;dg}} \tag{8}$$

The fact that equation (8) must hold for *all* $G\subset W$ and $G'\subset A$ is a rather strong testable implication of this model. Setting $G=W$ and $G'=A$ we have

$$\frac{\beta\left(A\right)}{\beta\left(W\right)}=\frac{\frac{\int_W \pi(\mathbf{a}_g,g)a_g^*\;dg}{\int_W a_g^*\;dg}}{\frac{\int_A \pi(\mathbf{a}_g,g)a_g^*\;dg}{\int_A a_g^*\;dg}} \tag{9}$$

[13]To take the ratio we assume that $\int_G a_g dg$ and $\int_{G'} a_g dg$ are positive, that is, that the treator searches both G and G'. We will return to this point in the next section.

The left-hand side, if different from 1, represent (relative) bias. The right-hand side is the ratio of the average profitability in the two racial groups. The idea that bias can be detected by comparing the profitability across subgroups is usually attributed to Gary Becker, who observed that a firm which discriminates against minority employees uses labor inputs less efficiently, and therefore should have lower profits, than a non-discriminating firm.

Proposition 1 *Suppose a treator solves problem (5). Then* $\frac{\beta(A)}{\beta(W)}$ *is equal to the average profitability in race* W *divided by average profitability in race* A*.*

A very useful feature of this proposition is that it is possible to ascertain bias even in the presence of "productive" unobservables which may be correlated with race.

4.2 Extension: General equilibrium

A prediction of this model is that only the groups with the highest value of $\beta(g)\,\pi(\mathbf{a}_g, g)$ are treated. If the functions $\beta(g)$ and $\pi(\mathbf{a}_g, g)$ were chosen arbitrarily, we would typically expect just one group to attain the highest value of $\beta(g)\,\pi(\mathbf{a}_g, g)$, and therefore only one group to be treated.[14] This would be an unrealistic implication of the model.[15] But in fact this implication need not follow if the value of $\pi(\mathbf{a}_g, g)$ is determined in equilibrium through the dependence on $\mathbf{a}_g$. Positing a dependence of $\pi(\mathbf{a}_g, g)$ on $\mathbf{a}_g$ is reasonable in applications in which the treated agents react to the behavior of a mass of treators whose actions generate the aggregate action $\mathbf{a}_g$. For example, in enforcement situations citizens in group g may

[14] More precisely, this unrealistic implication would hold provided that all resources can be absorbed by treating the "highest scoring" group.

[15] And, in addition, either the numerator or the denominator of equation (7) would be ill-defined.

decrease their rate of carrying contraband if they expect to be the focus of enforcement by the entire police department. To make this argument formal, we now sketch a bare-bones model that incorporates the response of the treated to treatment.

By assumption, the impact of a single treator on $\mathbf{a}_g$ is nil, so we need to model a mass of treators, the actions of which give rise to the aggregate action $\mathbf{a}_g$. We consider a rather stark model in which there is a mass μ of treators which are identical in all respects to the treator described in Section 4.1. The problem is now a general equilibrium problem involving a mass of treators. The equilibrium is described by the following set of conditions:

$$a_g^* \in \arg\max_{\{a_g\}} \int \beta(g)\, \pi\left(\mathbf{a}_g^*, g\right) a_g \; dg \quad \text{s.t.} \quad \int a_g \; dg \leq C$$

$$a_g \geq 0 \text{ for all } g$$

$$\mathbf{a}_g^* = \mu \cdot a_g^*.$$

The optimization problem is the same as that in Section 4.1 except for the last line. The last line is an accounting equation specifying that the parameter $\mathbf{a}_g^*$ is the aggregate of the individual actions a_g^* of all treators. Note that this definition of equilibrium requires that all treators take the same actions.[16]

This general equilibrium model differs from the previous one in that now the value of the function $\pi(\mathbf{a}_g, g)$ is determined as part of the equilibrium. If the response of $\pi(\mathbf{a}_g, g)$ to variations in $\mathbf{a}_g$ is sharp enough, then we can expect many groups to yield exactly the same

[16]This is with little loss of generality, in the sense that, while there are a multiplicity of other equilibria in which different treators take different actions, the set of values $\pi\left(\mathbf{a}_g^*, g\right)$ is the same across all these equilibria.

value of $\beta(g)\,\pi(\mathbf{a}_g, g)$ in equilibrium. The intuition for this equalization is that, if only one group was treated in equilibrium, then its members would respond by decreasing its π. If this response is sharp enough, the profitability of that group would fall below that of other groups. But then it is optimal to treat those groups too. Hence in equilibrium more than one group would be treated.

Formally, a sufficient condition for more than one group to be treated is that for every g there exists a g' such that

$$\pi(\mu C, g) < \pi(0, g').$$

A sufficient condition for all groups to be treated is an Inada condition of the form

$$\pi(0, g) = \infty \text{ for all } g.$$

Regardless of whether these conditions hold, Proposition 1 carries over to this setting. Therefore, the test to detect bias is unchanged when we go to the general equilibrium model.

The general equilibrium version of the model ties up two loose ends. First, as mentioned above, it explains why we should not be surprised that several classes can be treated simultaneously in equilibrium. Second, it pins down the intensity of treatment across treated groups, which would be indeterminate in the partial equilibrium model. In the general equilibrium formulation, although individual treators are indifferent between any allocation of their treatment across the groups that receive positive treatment in equilibrium, the aggregate level $\mathbf{a}_g^*$ is uniquely determined in equilibrium. Indeed, in equilibrium the vector $\mathbf{a}_g^*$ is set so as to equalize $\beta(g)\,\pi\left(\mathbf{a}_g^*, g\right)$ across all groups that receive positive treatment in equilibrium. Thus, in equilibrium disparities in treatment $\mathbf{a}_g^*$ across groups or classes of groups can arise even if $\beta(A) = \beta(W)$, provided that $\pi(\cdot, g) \neq \pi(\cdot, g')$ for some $g \subset A, g' \subset W$. In other words, differences in the intensity of treatment partially reflect (observable or unobservable) differences in the reaction of the treated to treatment. This is why information about the intensity of treatment across groups is generally not sufficient to identify bias.

A more general version of this "general equilibrium" setting is the one where we allow the treated to respond not only by reducing their crime, but also by disguising themselves as members of other groups. In the case of the police, for example, if members of group g are policed very intensely, they have the option not only of decreasing their crime rate, but also of disguising themselves as members of group g'. In this setup, the return from treating group g would depend not only on the intensity with which group g is treated, but also on the intensity with which other groups are treated. If we denote by $[\mathbf{a}_{-g}]$ the vector of treatment intensities for all groups other than g, we can write

$$\pi(\mathbf{a}_g, [\mathbf{a}_{-g}], g).$$

The main point is that Proposition 1 carries over to this setting.

The identification strategy described in this section was used in Knowles *et. al.* (2001) and Persico and Todd (2006). In both cases, it was applied to vehicular searches,[17] and the function $\pi(\mathbf{a}_g, g)$ was taken to be the probability of finding contraband. In this environment the right-hand side of (9) corresponds to the fraction of searched motorists of group G who are found with contraband—the so-called "hit rates." Both papers found hit rates to be very

[17] On Maryland's I-95 in one case, in Wichita, Kan. in the other.

similar not only between African American and whites, but also along by a number of other characteristics (sex, time of day, age of the driver, etc.).[18] This equalization seems unlikely to happen randomly. A possible conclusion, therefore, is that motorists are responsive to search intensity as assumed in this model and that police officers are not biased against African Americans.

Ravina (2007) and Pope and Snydor (2008) study a decentralized lending market in which lenders set interest rates to compete for loans of fixed size. The intuition behind the theoretical setup developed by Pope and Snydor (2008) can be easily understood by studying a slightly different game, one in which lenders allocate money a_g to borrowers of type g subject to a budget constraint. The expected profit from one dollar lent to group g is given by the function $\pi(\mathbf{a}_g, g)$. As the aggregate amount $\mathbf{a}_g$ lent to that group increases, $\pi(\mathbf{a}_g, g)$ is assumed to decrease reflecting a decrease in the equilibrium interest rate. This formulation casts Pope and Snydor (2008) directly within the framework of this section. Pope and Snydor find that loans to African Americans produce a lower rate of return than loans to whites. This finding is consistent with some kind of market discrimination against whites, or with a failure by lenders to fully take into account what race signals about the probability of repayment.

[18] Hit rates on Hispanics are significantly lower in both data sets, possibly suggesting some measure of discrimination against them.

5 Identification using variation in the observability of race

Suppose we had access to exogenous variation in the observability of race. Suppose, that is, that we could observe the treator's behavior when he can observe race and when he cannot. Intuitively, the color-blind setup might seem the ultimate benchmark against which to compare ordinary (and thus possibly biased) behavior. This intuition is based on the notion that any disparities that arise when race becomes observable are due to bias. This intuition is valid if the treator observes no other variable other than race that can improve the targeting of the treatment, but is not valid otherwise. The next example shows that an unbiased police officer looking for contraband will stop more African Americans when race is not observable than when it is. This difference, by construction, cannot be ascribed to racial bias on the part of the officer. Rather, the difference arises from the specific pattern of correlation between race and some other variable observed by the officer.

Example 2 *An unbiased officer looking for contraband can stop and search 100 people. He can observe the color of their car (dark or light) and possibly the race of the driver. The table below shows the probability that each subgroup carries contraband / the numerosity of each subgroup.*

	african american	*white*
dark colored car	*0.5/50*	*0.4/50*
light colored car	*0/70*	*0.6/70*

If the officer can see the driver's race, then he will stop all 70 whites with light-colored cars *and 30 African Americans with dark-colored cars. If the officer cannot see race, his best bet is to select dark-colored car drivers. Therefore, the officer who can see race stops 30% African Americans, the color-blind officer stops 50% African Americans, and in both cases the officer is unbiased.*

The example demonstrates that, as race becomes observable, changes in treatment between the two races are driven by the correlation between race and other characteristics observed by the police (car color, in the example). Therefore, bias cannot be identified solely from changes in treatment that arise as race becomes observable.

Despite this theoretical point, in some cases variation in the observability of race may help shed light on the presence (or absence) of bias. Grogger and Ridgeway (2006) study how the fraction of black drivers stopped by the Oakland police varies between day-time and night-time. Presumably, the race of the driver is more difficult to observe at night, yet essentially no variation is detected in the fraction of black drivers stopped. I find the absence of variation rather illuminating.

We can adapt the same logic to situations such as blind v. non-blind musical auditions.

Example 3 *An orchestra conductor auditions 120 men and 120 women musicians for a total of 100 jobs. As these are tenured jobs, he wishes to select musician with the greatest future musical ability. He can observe the quality of their performance in the audition today (good or bad), and the musician's gender if the audition is not blind. The table below shows the probability that each gender$\times$audition quality pair ultimately becomes a good musician through their career, and the numerosity of each subgroup.*

	man	*woman*
good audition	*0.3/50*	*0.2/50*
bad audition	*0.25/70*	*0.1/70*

If the conductor can see the musician's gender, then he will hire all 100 men. If the conductor cannot see gender, his best bet is to select based on audition performance and he will hire 50 men and 50 women. Therefore, blind auditions lead to an increased percentage of females hired even though the conductor is unbiased.

Goldin and Rouse (2000) find that female musicians are more likely to be hired when the audition is blind. Although Example 3 cautions against using this finding to draw inference about bias, that example relies on the difference between performance in the audition and future musical ability (the outcome of interest). When these two are closely aligned, as it is plausible they might be in practice, then the evidence presented by Goldin and Rouse becomes highly suggestive of bias in the sense modeled in this paper.[19,20]

6 The inframarginality problem

The term "inframarginality" refers to a problem that clouds identification. To facilitate exposition, let's assume that $\pi(a_g, \mathbf{a}_g, g) = \pi(g) \cdot a_g$. We will relax this assumption in Section 6.1.

[19] The examples presented here are somewhat similar in flavor to Heckman's critique of pair audit studies (see Heckman 1998), even though the second does not explicitly rely on the presence of productive variables used by the treated and unobserved by the researcher.

[20] In addition, from a legal viewpoint the mere practice of using gender as a predictor of a musicians' musical abilities conditional on audition performance is probably illegal per se. This gives rise to what Ayres (2002) refers to as the "subgroup-validity problem."

We conceptualize the inframarginality problem by adding the following constraint to problem (5):

$$a_g \leq I(g) \text{ for all } g. \tag{10}$$

The interpretation is that no group can be treated with intensity exceeding $I(g)$. This constraint captures, albeit in a stylized way, situations in which there are frictions in the reallocation of resources across groups—in this case, sharp diseconomies of scale when the group g is treated with intensity exceeding $I(g)$. We define marginal and inframarginal groups as follows.

Definition 4 *A marginal group is a group for which the optimal treatment satisfies* $0 < a_g^* < I(g)$. *An inframarginal group is a group for which* $a_g^* = I(g)$.

In what applied situations is constraint (10) likely to bind for some group? Consider for example a physician who with probability a_g prescribes further testing to patients with symptoms g. In this case $I(g) = 1$ for all g, because a probability cannot exceed 1. The physician may choose to prescribe further testing for *all* patients who show sufficiently serious symptoms. For these groups of patients, the probability of being referred is 1, so the constraint (10) binds and they represents the inframarginal groups.

The Lagrangean for this more-constrained problem is the following:

$$\mathcal{L}\left(\{a_g\}, \lambda_0, \lambda_1(g)\right)$$
$$= \int \beta(g)\,\pi(g)\,a_g\ dg - \lambda_0\left[\int a_g\ dg - C\right] - \int \lambda_1(g)\left[a_g - I(g)\right]\ dg$$

subject to $a_g \geq 0$ for all g. If an optimal a_g^* is strictly positive and finite, then maximization of the Lagrangean implies that $[\beta(g)\,\pi(g) - \lambda_0 - \lambda_1(g)]$ is maximal and equal to zero. The presence of the term $\lambda_1(g)$ creates a problem for the identification strategy. The analogue of equation (7) now involves terms such as $\int \lambda_1(g)\,dg$, which means that the ratio of average profitabilities (the right hand side of (9)) no longer directly reflects the difference in the β's.

To pinpoint the source of the problem, note that if we could observe $\pi(g)$ for two marginal groups, one in each race, then we would be able to identify bias. Indeed, for a marginal group we have simultaneously

$$\beta(g)\,\pi(g) - \lambda_0 - \lambda_1(g) = 0 \text{ and} \tag{11}$$

$$\lambda_1(g) = 0. \tag{12}$$

The first equality reflects the fact that $a_g^* > 0$, the second reflects the fact that $a_g^* < I(g)$. Provided we have two marginal groups g_m^A and g_m^W, in race A and W respectively, we could use (11) and (12) to get

$$\frac{\pi\left(g_m^W\right)}{\pi\left(g_m^A\right)} = \frac{\beta(A)}{\beta(W)}. \tag{13}$$

Thus, if hypothetically we could observe the average profitabilities for both marginal groups, then we could read the bias off of the ratio in profitabilities. But, as mentioned above, we do not directly observe the profitability for the marginal groups because we cannot recognize marginal groups. Rather, we observe an aggregate of profitabilities $\int_G \pi(g)\, a_g dg$ over a broader set of groups which may include g_m, a marginal group. The confounding groups $g \neq g_m$ are called "inframarginal," and thus the identification problem is referred to as "inframarginality problem."[21]

6.1 Response from motorists alleviates the inframarginality problem

The inframarginality problem arises when at the optimal solution the constraint $a_g \leq I(g)$ is binding for at least one g. Obviously, the constraint is less likely to bind when $I(g)$ is large. But what other factors can alleviate the inframarginality problem? A notable such factor is the response of treated to the treatment. In this section we show that a general equilibrium model like the one presented in Section 4.2 indeed can attenuate the impact of the inframarginality problem. The general equilibrium problem is described by the following

[21] Dharmapala and Ross (2004) and Bjerk (2007) studies a model where police enforcement is subject to the inframarginality problem.

set of conditions:

$$a_g^* \in \arg\max_{\{a_g\}} \int \beta(g)\, \pi(\mathbf{a}_g, g)\, a_g \; dg \quad \text{s.t.} \quad \int a_g \; dg \leq C$$

$$a_g \geq 0 \text{ for all } g.$$

$$a_g \leq I(g) \text{ for all } g.$$

$$\mathbf{a}_g = \mu \cdot a_g^*$$

By assumption, the function $\pi(\mathbf{a}_g, g)$ is decreasing in $\mathbf{a}_g$. In the case of police searches, for example, if group g is searched more intensely, that group will reduce its illegal activities. Intuitively, we should expect this property to alleviate the inframarginality problem. This is because the inframarginality problem arises from the treator's desire to treat group g with intensity greater than $I(g)$. When that group is allowed to respond, as it is in this formulation of the problem, the profitability of treating group g will decrease, which will decrease the incentives for the treator to focus on group g in the first place. A polar case that brings this force into sharp relief is the case where

$$\pi(\mu I(g), g) = 0 \text{ for all } g. \tag{14}$$

This assumption means that treating group g becomes unprofitable before the intensity of its treatment hits the constraint $I(g)$. If this assumption holds, then in equilibrium the inframarginality constraint will not bind and the inframarginality problem does not arise.

6.2 Identification in the presence of inframarginality: Comparing the performance of two treators

Absent sources of exogenous variation in C, the literature has been able to deal with the identification problem only partially. The identification strategy has been to compare the performance of two different treators, labelled 1 and 2. From these differences in performance, it has been shown that it is sometimes possible to identify which of the two treators is *more* biased against members of race R. Formally, the identification strategy sometimes allows to reject the hypothesis that

$$\frac{\beta_1 (A)}{\beta_1 (W)} \geq \frac{\beta_2 (A)}{\beta_2 (W)} \tag{15}$$

The identification strategy does not, however, tell us whether $\beta_i (A) / \beta_i (W)$ exceeds 1, so we cannot exclude that both treators are biased in favor of whites, or possibly against whites.

Suppose we had two treators, $i = 1, 2$, with possibly different β_i's and C_i's, both of whom are treating the same population.[22] Assume that for $R = A, W$ each treator had a marginal group $g^R_{m,i}$. For each treator i we have, from equation (13),

$$\frac{\beta_i (A)}{\beta_i (W)} = \frac{\pi \left(g^W_{m,i} \right)}{\pi \left(g^A_{m,i} \right)}$$

Suppose (15) holds. Then

$$\frac{\pi \left(g^W_{m,1} \right)}{\pi \left(g^A_{m,1} \right)} \geq \frac{\pi \left(g^W_{m,2} \right)}{\pi \left(g^A_{m,2} \right)}.$$

[22] This is a very important assumption which sometimes needs to be defended in practical applications. For example, in some studies the treatment is a search (or a ticket), and the population is the motorists stopped by an officer. Clearly, in this case the assumption that two different officers stop the same population of cars needs to be defended.

This equation *excludes* the possibility that

$$\pi\left(g_{m,1}^{W}\right) < \pi\left(g_{m,2}^{W}\right) \ \& \ \pi\left(g_{m,1}^{A}\right) > \pi\left(g_{m,2}^{A}\right). \tag{16}$$

Equation (16) can be translated into a statement about the intensity of treatment. To see how, observe that by equation (11), a group in race R is treated by treator i if and only if $\pi(g)$ exceeds a lower bound which, by definition, is exactly $\pi\left(g_{m,i}^{R}\right)$. Therefore for both treators the number of searches in race R are given by the same decreasing function of $\pi\left(g_{m,i}^{R}\right)$. Then equation (16) implies that treator 1 searches more whites than treator 2, and treator 1 searches fewer African-Americans than treator 2. Equation (16) therefore demonstrates the following proposition, which is contained in Anwar and Fang (2006).

Proposition 5 *Suppose two treators solve problem (5) with treator-specific C_i and $\beta_i(\cdot)$, and with the additional inframarginality constraint (10). Suppose treator 1 treats more whites than treator 2 and, at the same time, treator 1 treats fewer African-Americans than treator 2. Then $\frac{\beta_1(A)}{\beta_1(W)} < \frac{\beta_2(A)}{\beta_2(W)}$.*

It is worth remarking that this proposition holds regardless of the values of the C_i's. Of course, if the C_i's are very different then we may be unlikely to observe the constellation of treatment intensities described in Proposition 5, and so we may not be able to rule out hypothesis (15).

A number of papers compare the treatment by different groups of treators. In the context of vehicular searches, Antonovics and Knight (2007) find that police are less likely to search motorists of their own race and interpret this finding as evidence of bias. Price and Wolfers

(2008) study the number of fouls in basketball games called by refereeing crews with varying minority compositions. Applied to their setup, $\pi(g)$ represents the severity of a foul with characteristics g (where g includes both foul and player characteristics), $a(g)$ represents the number of fouls of type g called, and $I(g)$ represents the number of fouls of type g committed in a typical game. The coefficient $\beta_i(R)$ multiplies the severity of the foul in the referee's payoff, so a high value of $\beta_i(R)$ means that crew i is biased against race R. Inspection of Table 3 in Price and Wolfers (2008) shows that an ordinal property like that described in Proposition 5 holds. Indeed, they find that majority white refereeing crews (treator 2 in our language) assess 4.330 fouls per game on black players and 4.954 fouls per game on white players, while majority black crews (treator 1 in our language) assess 4.329 fouls per game on black players and 5.023 fouls per game on white players. Therefore it is legitimate to conclude, within the model presented in this section, that majority-white refereeing crews are relatively more biased against African American players.[23]

Equation (16) also has implications for average profitability rates. The average profitability by race is given by

$$H(\underline{\pi}, R) = \int_{\substack{\pi(g) \geq \underline{\pi} \\ g \subset R}} \pi(g) \cdot \frac{I(g)}{\int_{\substack{\pi(g) \geq \underline{\pi} \\ g \subset R}} I(g)\ dg}\ dg$$

Since $H_R(\underline{\pi})$ is monotone increasing in $\underline{\pi}$, equation (16) implies the following proposition, also contained in Anwar and Fang (2006).

Proposition 6 *Suppose two treators solve problem (5) with treator-specific* C_i *and* $\beta_i(\cdot)$,

[23] Of note, the common practice of signing the difference-in-difference in the foul-per-game (0.70 > 0 in our case) represents a different, less restrictive test than the ordinal test in Proposition 5. For such a test to correctly be interpreted as identifying bias, more stringent restrictions need to be placed on the environment than are placed in this section.

and with the additional inframarginality constraint (10). Suppose treator 1 has a lower average profitability on whites than treator 2 and, at the same time, treator 1 has a higher average profitability on African-Americans than treator 2. Then $\frac{\beta_1(A)}{\beta_1(W)} < \frac{\beta_2(A)}{\beta_2(W)}$.

This proposition has been used by Anwar and Fang (2006) in the context of vehicular searches by the Florida Highway Patrol.[24] They find that hit rates on whites are higher than hit rates on African Americans, which in the framework of Section 4 would denote racial animus against African Americans. Yet they also find evidence suggesting that the appropriate model is closer to the one presented in Section 6. Therefore, they look to apply Proposition 6. They find that the hit rates vary by race of the officer, and yet the hit rates of white and black officers do not line up as posited in Proposition 6. Therefore, they argue that it is not possible to conclude that black and white officers are ranked in terms of bias. Based on more years of data from the same agency, Ilic (2008) re-does Anwar and Fang's (2006) calculations and finds that, aggregating over the entire period, hit rates are roughly equal across black and white motorists and by race of the officer making the search. On the whole, the more comprehensive data set analyzed by Ilic appears to be closer to the framework of Section 4. It is not clear at present what features of the data generate the discrepancy between the results of Anwar and Fang (2006) and Ilic (2008).

[24] Rowe (2008) derives a related test and applies it to ticketing probabilities.

7 Identification with non-atomistic treators

Until now we have dealt with atomistic treators, whose impact on aggregate quantities was assumed to be negligible. As such, a treator could devote as many resources as needed to group g without affecting its behavior, and so the treator's payoff function was linear in resources. When treators are large they may have an impact on aggregate quantities, and nonlinearities are likely to arise in the treator's objective function. In that case, the identification is highly sensitive to the precise objective of the treator. In what follows we shall concentrate on the case of a monopolistic treator,[25] so we may write without loss of generality

$$\pi\left(a_g, \mathbf{a}_g, g\right) = \pi\left(a_g, g\right).$$

We assume that $\pi\left(a_g, g\right)$ is concave in a_g. This apparently conventional assumption will be discussed in Section 8.

Let us dispose first of a particularly simple case. In certain environments, the legitimate objective of the treator might be to *equalize* an objective function across categories. For example, a judge may be required to set bail levels for different defendants so as to achieve a given "appropriate" (race-independent) level of probability of flight. In these environments, the treator's legitimate objective is to set a_g^* so as to achieve

$$\pi\left(a_g^*, g\right) = \overline{\pi} \text{ for all } g.$$

We can think of a biased treator in this setting as one whose value $\overline{\pi}$ comes to depend on

[25] We do not deal with the case of oligopolistic treators.

g. This conceptualization is formally analogous to condition (6) because the value of the "legitimate" portion of the objective function at the optimum is proportional to a factor that captures bias. Therefore, the identification strategy developed in Section 4 applies to this case.[26] Ayres and Waldfogel (1994) apply that strategy to look for racial bias in the judge's decision of the level at which to set bail.[27] Although they do not directly observe the probability of flight (corresponding to $\pi\left(a_g^*, g\right)$), they observe the fee charged by bail bondsmen to defendants who borrow to pay the bond. The assumption is that the size of the fee reflects flight probability.[28] They find that, compared to whites, African Americans are charged lower fees, which suggests a lower probability of flight and therefore "too large" a bond.[29]

A more challenging setup, and a fairly natural one, is that in which the treator solves problem (1) subject to constraints. We shall proceed under the assumption that the constraint is given by (4), just as in Section 4.1. Even so, the identification problem is qualitatively different from the one solved in that section, as we shall see. An example of our problem is that of a police chief allocating manpower across neighborhoods g with a legitimate goal of minimizing aggregate crime across all neighborhoods. Since it is reasonable to assume that the chief's actions affect crime rates in each neighborhood, the treator's objective function is likely

[26] We also assume implicitly that there is no inframarginality constraint.

[27] Ayres and Waldfogel (1994) must be credited for recognizing that the identification strategy presented in Section 4 is robust to unobservables.

[28] One might object to the assumption that the function transforming the bond fee into a flight probability is the same for blacks and whites. If blacks are poorer than whites, a lower bond fee might suffice to achieve a given flight probability.

[29] As well, they find that those charged with more severe offenses pay lower rates, which within the context of the model suggests that judges set bail so as to achieve a lower probability of flight for more severe offenses.

non-linear. The programming problem is now

$$\max_{\{a_g\}} \int \beta(g)\,\pi(a_g, g)\ \ dg \quad \text{s.t.} \quad \int a_g\ dg \leq C \tag{17}$$

$$a_g \geq 0 \text{ for all } g.$$

In the police chief's problem, the function $\pi(a_g, g)$ represents the negative of the crime rate in neighborhood g, and the coefficient $\beta(g)$ represents the weight given to neighborhood g's crime in the chief's objective function, so that a low $\beta(g)$ represents bias against neighborhood g. The associated Lagrangean is

$$\int \beta(g)\,\pi(a_g, g)\ \ dg - \lambda_0 \left[\int a_g\ dg - C \right],$$

subject to $a_g \geq 0$ for all g. The first order conditions are

$$\beta(g) \frac{\partial \pi(a_g, g)}{\partial a_g} - \lambda_0 = 0 \tag{18}$$

Suppose we have exogenous variation in resources C, and that we can compute how a marginal change in resources affects total crime in race R. That change is given by

$$\begin{aligned}
&\frac{d\int_R \pi\left(a_g, g\right) dg}{dC} \\
&= \int_R \frac{\partial \pi\left(a_g, g\right)}{\partial a_g} \frac{\partial a_g}{\partial C} dg \\
&= \int_R \frac{\lambda_0}{\beta\left(R\right)} \frac{\partial a_g}{\partial C} dg \\
&= \frac{\lambda_0}{\beta\left(R\right)} \frac{d\int_R a_g dg}{dC},
\end{aligned}$$

where the second equality follows from 18. We can therefore write

$$\frac{\lambda_0}{\beta\left(R\right)} = \frac{d\int_R \pi\left(g\right) a_g dg}{d\int_R a_g dg}.$$

The RHS represents the variation in total profitability detected in race R as a fraction of the change in treatment devoted to race R. Both terms can be recovered empirically.[30] It follows that

$$\frac{\beta\left(A\right)}{\beta\left(W\right)} = \frac{\frac{d\int_W \pi(g) a_g dg}{d\int_W a_g dg}}{\frac{d\int_A \pi(g) a_g dg}{d\int_A a_g dg}}.$$

Thus we have shown the following proposition.

Proposition 7 *Suppose a treators solves problem (17). Suppose we can observe the change in average profitability and a change in treatment due to exogenous variation in resources. Then* $\frac{\beta(A)}{\beta(W)}$ *is equal to the ratio of the changes in profitability in race* W *over race* A*, times the ratio of the changes in treatment in race* A *over race* W*.*

[30] If we think of the hit rate as an average effect, this is a marginal effect.

Of course, exogenous variation in C is not always available. Dominitz and Knowles (2006) provide parametric conditions under which no variation in C is required. Under their conditions, the success rates in the right-hand side of equation (9) provide information about bias.

8 Social optimality, equal treatment, and the absence of bias

In this section we move from identification to a more philosophical question. Within the framework analyzed in the previous question, what should be the goal of discrimination law?

The uncontroversial core goal of discrimination law is that the law (i) should aim at rooting out bias, (ii) with minimum disturbance to the economy; where (iii) bias is defined as disparate treatment of similarly situated individuals.[31] Why does this goal seem so reasonable? In part because our intuition is that removing "unwarranted" disparities moves us closer to the first best. More specifically, our intuition suggests the following statements should hold rather widely, and thus provide an "efficiency rationale" for discrimination law as it exists today.

a Bias (in the sense of taste for discrimination) is operationally equivalent to disparate treatment of similarly situated individuals.

[31] A subset of those who care about these issues, both in the public and in academia, would also be favorable to disturbing the economy, provided that the disturbance favors protected classes (minorities, women, etc.). This attitude is controversial, however, particularly among the non-protected classes.

b Disparate treatment of similarly situated individuals interferes with welfare maximization.

c Eliminating bias improves welfare.

d Interfering with unbiased treators moves the economy away from social optimum.

These four statements are valid sometimes–but not always. Statement c and d, for example, are correct by definition when there is a single treator and his "legitimate" objective function equals the social welfare function. However, there are many practical reasons why the legitimate objective function need not coincide with some reasonably agreed-to social objective function.[32,33] Statements a. and b. are also wrong sometimes, because they are based on a faulty intuition about how the optimal policy treats similarly situated agents. It is to this consideration that we turn first.

Example 8 (a, b) *(Persico 2002, Eeckhout et al. 2008) Consider two groups (A and W), each composed of 100 identical citizens. Every citizen will commit a crime unless he is policed with probability at least 49%. The police, acting as a monopolist, can police exactly 50 citizens and seeks to minimize total crime (regardless of the race of the criminal). If both racial groups are treated equally then each citizen has a probability 25% of being policed, and so all citizens will commit a crime. If the police focusses all its resources on one group, the W for example, then each member in that group will be policed with probability 50%, just enough to deter crime. Thus no citizen in group W, and all citizens in group A, will*

[32] One such reason is that in many practical applications the treator is an agent in a principal-agent relationship, and so the treator's "legitimate" objective function is designed at least in part to deal with the agent's incentive problem. We will return to this point later in this section.

[33] We choose to ignore henceforth one element of the social welfare function—the pleasure that the discriminators receive from discriminating. This omission does not drive our results.

commit a crime. Under the non-discriminatory strategy the crime rate is 100%, under the discriminatory one it is 50%. If we take crime minimization to be the social objective,[34] *then the discriminatory strategy is welfare-superior to equal treatment.*

The salient feature of this example is that the profit functions $\pi(\mathbf{a}_g, g)$, which in this case coincide with the welfare function (crime rate) are not concave in $\mathbf{a}_g$ (intensity of policing). It is this failure of concavity that gives rise to "optimal disparate treatment." This example demonstrates that even an unbiased social planner facing identical groups may want to treat these groups differently. Looking beyond this perfectly symmetric example, the more general point is that even if two groups are slightly different, the optimal solution may feature wildly disparate treatment.[35] In practice, then, this observation casts doubt on the expectation that conditioning on productivity should explain differences in treatment. That is not to say that race-based disparities are the unavoidable side-effect of optimization.[36] But Example 8 does show that the equivalence between the propositions "similarly situated individuals are treated differently" and "intent to discriminate" (in economic parlance, taste for discrimination), is not necessarily warranted.[37]

[34] We might want the welfare function to also account for the cost to the citizens of being policed. To the extent that the cost is the same for citizens of both groups, the aggregate cost of being policed is a constant in the welfare function and can therefore be ignored.

[35] Conversely, remedial policies resulting in a large impact on the allocation of treatment across groups may actually have a small effect on profitability.

[36] The previous example is somewhat artificial if we take the perspective that equal treatment by race is a value per se. In that case, one would presumably be able to segment the population into non-race based groups (say, by the initial of their last name), and implement the optimal policy based on those groups. In this way, we might be able to implement the optimal policy while reducing or eliminating the correlation between disparities in treatment and protected categories (race, gender, age, etc.).

[37] However, while at the optimal solutions the disparities may be correlated with a protected category, there are second-best allocations that in which the disparities need not be. Thus, if our goal is to achieve close-to-optimal and not-disparate by group allocations, the corrective action (legal, for example) had better look like a quota instead of relying on corrective action that alters the perceived productivity of individual groups. Quotas, of course, would be equivalent to segmenting treatment along arbitrary (but not protected) lines.

Having established that bias, in the sense of "taste for discrimination," cannot be conflated with "disparate treatment," let us now turn to an example in which eliminating bias is not welfare-improving. In this example points c. and d. do not hold. The ideas in this example are developed in Persico (2002), Alexeev and Leitzel (2004), and Persico and Todd (2005).

Example 9 (c, d) *Consider two groups (A and W), each composed of 100 citizens, each of whom will carry drugs unless the probability of being searched is sufficiently high. Citizens within each group are heterogeneous in their propensity to carry drugs, and so are deterred by different probabilities of being searched. In group W, exactly one citizen is deterred for every additional search applied to group W, whereas in group A it takes two additional searches to deter a citizen. There are 90 police officers who can each search exactly 1 citizen. Each officer chooses which group to search from in order to maximize the probability of a successful search. Suppose the police are unbiased. Then in equilibrium the police have to be indifferent between searching either group, and so the crime rate has to be equal in the two groups. This requires that* $100 - \mathbf{a}_W = 100 - (\mathbf{a}_A/2)$ *, which means that group A is searched twice as much as group* W*. In contrast, crime minimization requires directing all searches on group W, because it provides the highest marginal return to treatment. Making the police officers biased against* W *would move the equilibrium closer to the crime-minimizing allocation.*

In this example points c. and d. do not hold: making the officers biased against whites would improve welfare (reduce the crime rate), and interfering with unbiased treators (for example, forcing the unbiased police to search more whites) would improve welfare. The characteristic of the example is that, while the crime minimization (welfare maximization) problem is

$$\min_{\{\mathbf{a}_g\}} \int \pi\left(\mathbf{a}_g, g\right)\ dg,$$

(we take $\pi\left(\mathbf{a}_g, g\right)$ to be the crime rate), each police officer maximizes successful searches, solving

$$\max_{\{a_g\}} \int \pi\left(\mathbf{a}_g, g\right) a_g\ dg.$$

The two problems give rise to different first order conditions (given by equations (18) and (6), respectively), so it should be no surprise that the aggregate behavior of individual police officers is not welfare maximizing in this example. Still, the example highlights a broader issue: the problem of incentivizing individual treators (the police officers) to make a costly treatment (effort in searching motorists) whose impact on the aggregate outcome of interest (crime rate) cannot be measured reliably (in this case because it is small). In such instances, it is necessary to come up with incentive schemes which reward individual effort (rewarding successful searches), and those schemes need not be collinear with the social welfare function. Whenever these incentive problems arise we are in a second-best world, and so there is little reason to think that points c. and d. apply. How important are such frictions in practice? Economists tend to think that incentive problems such as this are ubiquitous.

In circumstances where a-d may fail, the shared consensus for the core goals of discrimination law cannot rest on a-d. When c and d fail, for instance, bias can increase social welfare. What, then, is the basis for the core principles of discrimination law? The considerations presented above challenge us, I think, to dig deeper into this question.[38]

[38] My reading of Harcourt (2004) is that he recognizes these concerns, and he weaves them into a proposed evidentiary procedure to evaluate when using race as a factor in discretionary searches by law enforcement would be constitutionally acceptable. According to Harcourt, the police should be challenged if any disparity

9 Open questions

On the front of police enforcement, a central concern in the broaded field of racial profiling, an important task would be to greatly extend the set of jurisdictions that are the subject of analysis. A difficult, but very interesting question would be to understand whether the models analyzed in this paper are applicable to other jurisdictions, and if not, what kind of alternative models need to be developed. Painting a broad picture of the phenomenon of racial disparities in enforcement is important because existing studies are limited in their scope. In this connection, it should be noted that enforcement agencies themselves frequently collect and analyze their own data. Therefore, in theory this task could be accomplished by the enforcement agencies themselves, if they were able to deal with identification and other data issues that necessarily arise in any practical situation. More realistically, the fact that enforcement agencies collect enforcement data means that the data exist in machine form and can, at least in principle, be requested by researchers via Freedom of Information Act requests.

On the identification front, a natural next step is to develop models of multi-stage treatment. In many contexts, the same treator treats an agent in several stages. In the labor context, for example, the employer first selects applicants and then retains them, promotes them, and pays them wages. Altonji and Pierret (2001) develop a multi-stage model of employment, and their results indicate that employers do not condition their wage on race at the first se-

by race is observed. If the police is unable to come up with other factors that eliminate the statistical effect of race on the search decision, then the police would then have to show: (i) that race is predictive of crime; and (ii) that the percentage of minorities among the criminals who happen to be searched be no greater than the percentage of minorities within the criminal population at large; and (iii) that the use of race helps decrease aggregate crime.

lection stage (wage at first hiring), suggesting that employers do not behave as predicted by the statistical discrimination model of Arrow (1973).[39] Barnes (2005) develops a statistical selection model dealing with vehicular stop and search data. She is able to provide information about some observable characteristics of the vehicle stopped and not searched (in our language, she is able to provide information about the g's for which treatment is equal to zero). By means of the statistical model she then infers the probability of carrying contraband of those g's who are not searched.[40] In general, we would expect that the modeling the interaction between multiple stages of decision-making would pose new challenges—and opportunities— for identification.

A largely unexplored question is that of identification when the objective function π is not separable within or across classes, so that treating two agents, one in group g and the other in group g', does not give rise to the sum of $\pi(g) + \pi(g')$ but to a more complicated function $\pi_i(g, g')$ which is possibly treator-specific. This seems like a challenging problem, yet it is an important one because arguably many employers have specific production processes and cultures etc. that do not fit the additive, employer-independent specification.

A question of some theoretical interest is whether the bias is conscious or unconscious. This

[39] See also Coate and Loury (1993).

[40] Roughly speaking, the statistical methodology is based on the assumption that, after appropriately contolling for observables, the hit rates on those non searched are approximately equal to those of the searched. Since in her data hit rates are largely constant across races, the implication is that it would be possible to search some more whites without a decrease in hit rates. Within the framework put forth in the present paper, there are two potential issues with this procedure. First, even if motorists do not react to policing, and thus $\pi(\mathbf{a}_g, g) = \pi(g)$, if the police is made to search new $\widehat{g}$'s (more whites), it will necessarily be the case that the new groups searched have a return $\pi(\widehat{g})$ which is no higher, and possibly lower, than those g's who were being searched already. Second, if motorists do react to policing, and thus $\pi(\mathbf{a}_g, g)$ does depend on $\mathbf{a}_g$, then increasing $\mathbf{a}_g$ decreases $\pi(\mathbf{a}_g, g)$. For both reasons, we would expected that inducing the police to search more whites would bring down the average success rate on whites, particularly on those not previously searched. Despite these observations, Barnes (2005) adds value because it focusses on integrating the stop and search decisions and tackling the associated selection problems.

question appears to be challenging, both at an interpretive level and at an identification level. With an intellectual leap, we might phrase this question in our model as follows: is there bias in the $\beta(\cdot)$ functions (conscious bias), or are the treators *misperceiving* the expected profit function $\pi(\cdot)$, due perhaps to a failure to properly update (unconscious bias)?[41] On a related point, there is growing evidence suggesting that rapid decisions are more subject to (possibly) bias than more deliberate decision processes. As far as I know, discrimination law does not differentiate between "conscious" and "unconscious" intent to discriminate, which is somewhat interesting given the central role that intentionality and "mens rea" play in the legal system. It would be interesting to introduce these considerations into our analysis.

The analysis in the previous sections, including Section 8, suggests that the bias that discrimination law attempts to correct manifests itself differently in different environments. Specific disparities may indicate bias in certain environments but not in others. This suggests that evidentiary rules should be carefully tailored to specific areas of the law. For example, in the case of police enforcement, comparing the results from Sections 4 and 7 suggests that we should look for evidence of bias in different ways depending on whether we are concerned with the bias of the individual police officer(s) in the allocation of their discretionary searches, or whether we are concerned with bias in the allocation of aggregate resources by a police chief. Of course, evidentiary rules do differ across areas in US law. It would be interesting to assess the degree to which this variation can be explained as solving the kind of identification problems described in this paper.

[41] Bunzel and Marcoul (2008) present a theoretical model of improper updating in the context of police enforcement. In their setting, overconfindence in one's own abilities leads the police officers, over time, to focus enforcement disproportionally on one racial group.

On a more theoretical level, the analysis in Section 8 raises some normative question about the current goal of discrimination law—namely, to eliminate bias. As we have seen in Example 9, bias can sometimes help achieve social welfare due to a second-best logic. What should the law prescribe in these cases? Should we introduce race-based incentive schemes for police officers, for example, in order to improve social welfare? The natural reaction is to discount such instances as theoretical curiosities and therefore not worry about them. Of course, one might take the opposite viewpoint and use this argument to provide an efficiency rationale for non-neutral policies such as affirmative action.[42] The fact of the matter is that the adjudicator in a court of law is ill-positioned to assess questions of social optimality. This "ignorance argument" could be taken as a (not overly strong) argument in favor of the current goal of discrimination law.[43]

10 Conclusions

Discrimination is alleged along many lines—race, gender, age, disability, etc. Typically, the allegations arise in conjunction with disparities in outcomes. A disparity in outcomes, however, is not *per se* illegal, as it may reflect a correlation with (possibly unobserved) productive characteristics. What is illegal is intent to discriminate. Therefore it is important, both from a legal standpoint and from an intellectual standpoint, to be able to distinguish productivity-related (and thus justifiable) disparities from those reflecting discriminatory

[42] Of course, even in a second-best world, affirmative action need not necessarily be welfare-improving. On this point, see Coate and Loury (1993).

[43] Manski (2006) deals with the problem of a non-atomistic (indeed, a monopolistic) treator who seeks to maximize a profit function $\pi(a_g, g)$ with limited information about the shape of $\pi(a_g, g)$.

intent. Making this distinction using statistical data is generally seen as difficult, partly because discriminatory intent is viewed as a state of mind and therefore difficult to ascertain. In this paper we laid out a model in which to study the identification of a bias parameter. The model organizes several results that have been obtained in the literature in various applied contexts. The first message is that no single identification strategy works in every situation. Depending on the specific features of the problem, and on the variation that is available, different statistics represent valid evidence of bias. The second message is that identifying bias is not hopeless, and in fact many of the methods we have discussed have been successfully applied to real world data. Therefore, it is hoped that the body of work reviewed in this article will be of practical use.

References

[1] Alexeev, Michael and Jim Leitzel (2004) “Racial Profiling.” Manuscript, University of Chicago.

[2] Altonji, Joseph G., and Charles R. Pierret. (2001). “Employer Learning and Statistical Discrimination.” *Quarterly Journal of Economics* 116: 313-50.

[3] Antonovics, Kate L. and Brian G. Knight (2007) “A New Look at Racial Profiling: Evidence from the Boston Police Department.” Manuscript, October 10, 2007. Forthcoming, *Review of Economics and Statistics*.

[4] Amemiya (1985) *Advanced Econometrics*. Harvard University Press.

[5] Anwar, Shamena and Hanming Fang (2006). "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," *American Economic Review* 96(1): 127-151.

[6] Arrow, K.J. (1973) *The theory of discrimination*. In: O. Ashenfelter and A. Rees, Editors, Discrimination in Labor Markets, Princeton University Press, Princeton, NJ, pp. 3–33.

[7] Ayres, Ian (2002) “Outcome Tests of Racial Disparities in Police Practices.” *Justice Research and Policy* 4:131-142.

[8] Ayres, Ian, and Joel Waldfogel (1994). “A Market Test for Race Discrimination in Bail Setting.” *Stanford Law Review* 46(5): 987-1047.

[9] Barnes, Katherine Y. (2005). "Assessing the Counterfactual: The Efficacy of Drug Interdiction. Absent Racial Profiling, *Duke Law Journal* 54: 1089-1141.

[10] Becker, Gary S.(1971) *The Economics of Discrimination*. 2nd edition, University of Chicago Press.

[11] Bjerk, David (2007) "Racial Profiling, Statistical Discrimination, and the Effect of a Colorblind Policy on the Crime Rate," Journal of Public Economic Theory 9(3): 543-567.

[12] Bunzel, Helle and Philippe Marcoul (2008) "Can Racially Unbiased Police Perpetuate Long-Run Discrimination?" *Journal of Economic Behavior and Organization* , forthcoming.

[13] Coate, S. and G.C. Loury, (1993)," Will affirmative action policies eliminate negative stereotypes?". *American Economic Review* 83: 1220–1240.

[14] Dharmapala, D. and Ross, S. (2004). "Racial bias in motor vehicle searches: additional theory and evidence", *Contributions to Economic Analysis & Policy* 3(1), Article 12, http://www.bepress.com/bejeap/contributions/vol3/iss1/art12.

[15] Dominitz, Jeff (2003) "How Do the Laws of Probability Constrain Legislative and Judicial Efforts to Stop Racial Profiling?" *American Law and Economics Review* 5(2): 412-432.

[16] Dominitz, Jeff and John Knowles (2006) "Crime minimisation and racial bias: what can we learn from police search data?" *The Economic Journal*, 116: F368-F384.

[17] Eeckhout,Jan Persico, Nicola, and Petra Todd (2008) “Estimating Deterrence Effects Using Random Crackdowns: Theory and Evidence” Manuscript, New York University.

[18] Gelman, Andrew Jeffrey Fagan, and Alex Kiss (2007). “An Analysis of the New York City Police Department’s “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias.” Manuscript, Columbia University.

[19] Goldin, Claudia and Cecilia Rouse (2000). “Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians” *American Economic Review* 90(4): 715-741.

[20] Grogger, Jeffrey and Ridgeway, Greg (2006) “Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness.” J*ournal of the American Statistical Association* 101(475): 878-887.

[21] Harcourt, Bernard E. (2004) “Rethinking Racial Profiling: A Critique of the Economics, Civil Liberties, and Constitutional Literature, and of Criminal Profiling More Generally.” *University of Chicago Law Review* 71(4): 1275–1381.

[22] Heckman, James J. (1998) “Detecting Discrimination” *The Journal of Economic Perspectives* 12(2): 101-116.

[23] Heckman, James J., and Peter Siegelman (1993). “The Urban Institute Audit Studies: Their Methods and Findings.” In *Clear and Convincing Evidence: Measurement of Discrimination in America*, Michael Fix Raymond Struyk. Washington , D. C. : Urban Institute Press.

[24] Hernández-Murillo, Rubén and John Knowles (2004) "Racial Profiling or Racist Policing? Testing in Aggregated Data." *International Economic Review* 45(3): 959-989

[25] Ilic, Dragan (2008) "Racial Profiling, Prejudice, and Statistical Discrimination." Manuscript, University of Basel, 2008.

[26] Knowles, John, Persico, Nicola, and Todd, Petra. (2001). "Racial bias in motor vehicle searches: theory and evidence." *Journal of Political Economy* 109(1): 203–229.

[27] Manski Charles F. (2006) "Search Profiling with Partial Knowledge of Deterrence." *The Economic Journal* 116: F385–F401.

[28] Persico, Nicola (2002) "Racial Profiling, Fairness, and Effectiveness of Policing" *American Economic Review* 92(5): 1472-97.

[29] Persico, Nicola and Petra Todd (2005) "Passenger Profiling, Imperfect Screening, and Airport Security." *American Economic Review Papers and Proceedings* 95(2): 127-131.

[30] Persico, Nicola and Petra Todd (2006) "Using Hit Rates to Test for Racial Bias in Law Enforcement: Vehicle Searches in Wichita." *The Economic Journal* 116 : F351-F367.

[31] Price, Joe and Justin Wolfers (2008). "Racial Discrimination Among NBA Referees." NBER Working Paper #13206.

[32] Pope, Devin G. and Justin R. Sydnor (2008) "What's in a Picture? Evidence of Discrimination from Prosper.com" Manuscript, University of Pennsylvania, June, 2008.

[33] Ravina, Enrichetta (2007) "Beauty, Personal Characteristics and Trust in Credit Markets" Manuscript, New York University, December 2007.

[34] Rowe, Brian (2008) “Gender Bias in the Enforcement of Traffic Laws: Evidence based on a new empirical test” Manuscript, University of Michigan, July 2008.

[35] Todd, Petra (2006). “Testing for Racial Bias in Law Enforcement.” *Palgrave Dictionary of Economics.*