A Behavioral Theory of Discrimination in Policing[*]

Ryan Hübert[†] Andrew T. Little[‡]

March 2, 2020

Abstract

Racial disparities in policing are well documented. In addition to officer animus towards some groups ("taste-based discrimination"), these could be driven by officers' beliefs that crime rates are higher in some communities ("statistical discrimination"). But where do these beliefs come from, and what if they are incorrect? We analyze a formal model where officers form beliefs using crime statistics, but make a common inferential mistake by not fully adjusting for the fact that they will detect more crime in more heavily policed communities. This creates a feedback loop where officers (incorrectly) believe there is more crime in communities that are policed intensely, leading to persistent over-policing. We also show that discrimination driven by false beliefs is contagious across officers. This means that inferential mistakes can exacerbate discrimination even among officers with no animus and who sincerely believe disparities are driven by real differences in crime rates.

Keywords: policing, discrimination, formal theory, behavioral models

Word count: 11,960

[*] Authors are listed in alphabetical order. Many thanks to Katie Angell, Tony Carey, Tom Clark, Brandon Davis, Peter Enns, Jack Glaser, Becca Goldstein, Anna Harvey, Kris Kanthak, Brad Leveck, Emily Nix, Kyle Peyton, Tessa Provins, Gwen Prowse, Marcel Roman, Will Spaniel, Allison Stashko, Jeff Staton, Emily West, Michelle Wier, Jon Woon, and audience members at the 2019 Political Economy and Public Law Conference, APSA 2019, SPSA 2020, Berkeley, Stanford, Wash U, Pitt and USC for helpful comments.

[†] Assistant Professor, Department of Political Science, University of California, Davis. *Email*: rhubert@ucdavis.edu. *Website*: http://www.ryanhubert.com/. Corresponding author.

[‡] Assistant Professor, Department of Political Science, University of California, Berkeley. *Email*: andrew.little@berkeley.edu. *Website*: http://www.andrewtlittle.com/.

Police departments in the U.S. have become more professionalized over the past half-century, aiming to reduce arbitrary and abusive policing practices. And yet, dramatic racial disparities in policing persist. For example, 80% of people stopped under New York City's now-defunct "stop and frisk" policy were either black or Latino despite the fact that those two groups make up only half of the city's population (Goel, Rao, and Shroff 2016). In Boston, blacks comprised 63% of police stops that did not end in arrest from 2007 to 2010, even though only 24% of the population is black (The Sentencing Project 2015). Non-white motorists are more likely to be stopped than white motorists (Epp, Maynard-Moody, and Haider-Markel 2014).

There are two standard theoretical explanations for these disparities, one driven by preferences and one driven by beliefs. In a purely preference-driven account—often called taste-based discrimination—officers intrinsically like being punitive towards some groups, or dislike being punitive towards others. The second explanation—typically called statistical discrimination—is that there are real differences in the rates of criminal behavior across groups. Knowing this, police allocate more time policing members of groups with higher crime rates, or at least in geographical areas where those groups are concentrated.

Another explanation for policing disparities sits uncomfortably between these two standard explanations. What if officers police a certain group more intensely because they believe that the group has a relatively high crime rate, but this belief is incorrect, or at least exaggerated? In a proximate sense, this is discrimination driven by beliefs. But we might suspect that such inaccurate beliefs are more likely to be held by those with an intrinsic dislike of the group. If so, it may not make sense to think of the belief and preference channels as distinct and separable causes of discrimination since they could affect each other.

This is not just a hypothetical. Even highly-trained researchers make mistakes in interpreting crime data (Knox, Lowe, and Mummolo 2019; Knox and Mummolo 2020), and there is little reason to think that the relevant decision-makers will generally do better (see, e.g., Glaser 2015, for an overview). Even though law enforcement has become more data-driven in recent decades, police officials typically need to make decisions under time pressure without the benefit of the kind of statistical expertise that would enable high quality assessments about crime across communities. Indeed, the shift toward data-driven policing has been controversial. If departments rely on either faulty analysis or bad data, this can perpetuate disparities (see, for example, Harcourt 2007; Lum and Isaac 2016). Even the federal courts have weighed in to criticize flawed data analysis by police (e.g., *Floyd v. New York*, 959 F. Supp. 2d 540, S.D.N.Y. 2013).

We develop a set of models which allows for incorrect beliefs about the prevalence of crime across groups. However, we do not allow for *any* arbitrarily incorrect beliefs, and instead explicitly model how and why beliefs become incorrect via a specific inferential mistake. Officers in our models form beliefs about the relative prevalence of crime among members of different social groups based on the number of crimes the police detect, without fully accounting for the fact that more crimes are detected among members of groups that are policed more intensely (see Glaser 2006; 2015, for previous discussions of this mechanism, to which we contrast our approach and contribution in more detail below). We call this *non-conditioning bias*.

Our models also allow for police officers to have racial animus, and for crime rates to be different across groups. In the special case where officers form correct beliefs, these two mechanisms independently affect policing disparities, as in the standard accounts. However, once officers exhibit any non-conditioning bias, this creates a feedback channel where groups who are policed more intensely are viewed as having higher crime rates than they really do. This feedback loop amplifies whatever policing disparities would exist in the absence of non-conditioning bias. Put another way, a taste for discrimination causes inaccurate statistical discrimination.

We first formalize this argument in a simple and analytically tractable model with just one officer. Next, we extend the analysis to include multiple officers. When each officer has correct beliefs (as in standard models), all of their policing decisions are independent and neither officer's preferences or behavior affect the other. However, if one officer has a non-conditioning bias, then their beliefs will be influenced by the behavior of other officers. As a result, the discriminatory behavior of one officer can spill over and cause others to discriminate too.

A straightforward implication of the theory is that faulty data analysis by police departments may unwittingly exacerbate disparities. For example, if departments use data-driven algorithms to predict where crime is likely to occur (Collins 2018), the predictions generated by these algorithms may be highly discriminatory if they are based on simple counts of prior crimes detected by police.

Our model also suggests that policy responses to discriminatory policing should focus—at least in part—on alleviating distortions caused by non-conditioning bias. Most obviously, equipping decision makers in police departments with appropriate statistical training could help them avoid making faulty inferences from crime statistics. That said, success of a policy like this requires specific individuals to consistently and correctly apply this training even in the face of other professional (or even political) pressures. A more institutionally oriented policy response would focus on ensuring that policing decisions are not endogenous to the data generated by those decisions. For example, departments could establish fully independent crime analysis divisions that are barred from using data generated from policing, such as arrests.[1] Finally, if accurate analysis of crime data remains difficult or infeasible, forcing departments to allocate attention more evenly across communities in the short term—even if it seems less effective—can help department decision makers to form more accurate assessments of relative crime across all communities.

1 Explaining Policing Disparities

Policing disparities have been well documented.[2] And, there is convincing evidence that statistical estimates of policing disparities may actually understate the extent of those disparities (Knox,

[1] Note that our model suggests this separation could be useful even absent other concerns about departments' incentives to misreport data.

[2] For a list of studies documenting racial disparities in the criminal justice system (including in policing), see Balko (2018).

Lowe, and Mummolo 2019; Knox and Mummolo 2020). Disparities in criminal justice also have important social consequences. For example, they may reduce political participation (Weaver and Lerman 2010; Komisarchik, Sen, and Velez 2019)[3] and increase overpoliced populations' contact with the criminal justice system (Glaser 2015).

While the empirical question about whether policing disparities exist is mostly settled, the theoretical question about why these disparities exist is not. There is a robust literature devoted to cataloguing and teasing out the explanations for racial disparities in policing, as well as in the labor market and politics (for example, see Knowles, Persico, and Todd 2001; Anwar and Fang 2006; Persico 2009; Doleac and Stein 2013; Ewens, Tomlin, and Wang 2014; Butler and Broockman 2011; Broockman and Soltas 2018; Nathan and White, Forthcoming; Ash, Fagan, and Harris, Forthcoming).

Identifying the underlying causes of disparities is not only an academic exercise; there are important legal and policy implications. Under U.S. law, it is typically impermissible for the government (including police departments) to discriminate on the basis of membership in a protected class, such as race, gender, religion or national origin (see *Floyd v. New York*, 959 F. Supp. 2d 540, S.D.N.Y. 2013). However, in light of geographical patterns in both residential segregation and crime, policing disparities may emerge even when police departments use "facially neutral" (and legal) policing practices, such as deploying resources to high crime locations.[4] For policymakers and police departments seeking to reduce policing disparities, it matters why those disparities ex-

[3] However, Walker (2020) finds that "proximal contact" with criminal justice system (e.g., via a relative) is associated with increased political participation, and Peyton, Sierra-Arévalo, and Rand (2019) find certain kinds of positive, non-enforcement policing actually increase willingness to cooperate with police.

[4] For example, in *Floyd*, Judge Scheindlin notes: "I recognize that the police will deploy their limited resources to high crime areas. This benefits the communities where the need for policing is greatest."

ist. Different root causes call for different responses, from firing prejudiced officers and conducting bias training to changing policing tactics and reducing enforcement activities for certain kinds of crimes.

There are two "standard" explanations for disparities, both of which our model captures. The first has its origins in the theory of discrimination articulated by Becker (1957). According to this explanation, if police officers have animus toward some groups, this may directly influence where they want to focus their efforts. We capture this in our models by allowing for the possibility that police officers get higher marginal utility for detecting crimes among one group.

The other standard explanation for disparities emerges from the fact that group identity may be informative about crime. This is the mechanism driving the models of statistical discrimination that have emerged from the seminal work by Phelps (1972) and Arrow (1973). For example, if crime rates are different across groups, then given certain kinds of policing objectives (like reducing crime), it would be "rational" to police those groups with different intensities. A simple version of this argument is *police officers pull over non-white drivers at higher rates than white drivers because they believe that non-white drivers are more likely to be carrying contraband* (see Knowles, Persico, and Todd 2001; Anwar and Fang 2006). In this account, police officers may have no racial animus, but are simply trying to maximize the amount of contraband they detect using the information available to them.

Of course, the key question for this explanation is why a decision-maker has those beliefs in the first place. In standard models, an individual's beliefs must be correct. That does not imply that they cannot face any uncertainty. However, they must correctly assess this uncertainty in light of their experiences and environment. For example, an officer can only believe non-white motorists are more likely to carry contraband if the information they possess—processed according to Bayes' rule—indicates that non-white motorists in fact are more likely to carry contraband.

The typical requirement that beliefs be correct does not reflect misguided adherence to a fictional view about how humans think. If we let beliefs be incorrect in arbitrary ways, nearly any behavior can be trivially explained by assuming decision-makers have a false belief about what choices are in their best interest. Moreover, the requirement of correct beliefs—while stringent—has generated many new insights about the persistence of inequality between groups. A core contention of much of this research is that statistical reasoning of this kind can generate perverse outcomes with self-reinforcing and discriminatory stereotypes. Coate and Loury (1993) provide an early account of racial discrimination in labor markets, and similar findings explain racial disparities in both robberies and homicide (O'Flaherty and Sethi 2019), as well as the persistence of social segregation (Chaudhuri and Sethi 2008). More generally, the extent to which "top brass" within policing institutions are able to accurately assess the effectiveness of policing practices is a key concern and a potential explanation for policing disparities (see McCall 2019).

While models with arbitrary beliefs typically provide little analytic traction, assuming perfectly correct beliefs limits the range of formal theories of discrimination in a way that does not accord with many scholars' intuitions about how beliefs fuel discrimination. Bohren et al. (2019) survey the broader economics literature on discrimination and find that few papers consider, let alone test for, the possibility that disparities due to statistical reasoning may be based on biased or otherwise incorrect beliefs.

Our main technical contribution is to provide a formal theory of policing that allows for some limits on cognition in complex environments but that does not abandon the core contention that beliefs should be generated systematically from experiences. Our work is therefore tied closely to a large literature in behavioral economics focused on the ways in which decision makers deviate from standard assumptions in formal analysis (see Dhami 2016). Formal models of other political phenomena are increasingly exploring the implications of non-standard belief formation (for example, Levy and Razin 2015; Minozzi 2013; Acharya, Blackwell, and Sen 2018; Ogden 2019; Patty and Weber 2007).

More specifically, we build on work which examines how decision-makers may overweight or underweight information from sources like their prior beliefs (Kahneman and Tversky 1973;

Camerer 1995) or the decisions made by others (Eyster and Rabin 2005). Our formulation highlights a version of this that we believe to be particularly relevant to policing: decision makers' limited ability to condition on all relevant information—specifically policing intensity—when making inferences about crime rates. We call this *non-conditioning bias*.

Of course, we are not the first to suggest that cognitive biases are important for understanding policing. In fact, Eckhouse (2019) argues that work on bias in policing overemphasizes cognitive biases relative to structural factors which put police disproportionately in contact with certain communities, with evidence from a change in the stop-and-frisk policy in New York City. Our model highlights that cognitive and structural biases are not necessarily competing explanations, as they can be mutually reinforcing sources of disparities.

Closest to our argument, Glaser (2015) argues that racial disparities in policing are fueled by a feedback loop driven by cognitive biases. Our model complements and builds on this work in several ways. First, by explicitly formalizing the officer's decisions and beliefs, we can make more precise predictions about how factors like racial animus, real crime rates, and the severity of our behavioral bias affect these outcomes.[5] Second, we show how discrimination can become contagious and spread to other officers who may not have discriminated if they were acting on their own. And finally, we make a broader contribution to the literature on discrimination by highlighting that taste-based and statistical discrimination are not distinctly separate explanations for disparities, except under the extreme, knife-edged assumption that officers have correct beliefs.

Finally, a broader intervention we aim to make in the study of discrimination and prejudice is to point out that this work frequently blurs the line between preferences and beliefs as drivers of disparities. Canonical work classifies it as prejudice when a person holds an inaccurate belief that members of a certain group disproportionately engage in undesirable behaviors (e.g., commit

[5] Glaser (2006) also contains a formal model of the implications of unequal policing allocations on incarceration and the efficiency of catching crime. Our focus is on the causes of unequal allocations.

more crimes) or have undesirable traits (e.g., are more "criminal").[6] In contrast, a key feature of our model is that we clearly distinguish the potential mechanisms driving disparities—preferences and beliefs—and then demonstrate how animus toward a group (preferences) can *endogenously cause* overly negative and incorrect beliefs about that group via non-conditioning bias.

2 Model of a Unitary Officer

We start with a model of a single police officer (pronoun "he"), who we primarily interpret as a high level official who makes decisions for the department as a whole, such as the chief of police. The officer makes a choice about how to allocate resources toward policing two groups, A and B. While we do not introduce notation for the group size, the model is easiest to interpret as one where the two groups are equally numerous.[7]

The officer has a unit of resources, which we primarily interpret as time, to allocate between policing the two groups. Let w_A represent the share of time spent policing group A, with $w_B = 1 - w_A$ left for group B. We assume that the officer can choose to allocate his time evenly between the two groups, but can also choose to police on group more than the other. However, the officer

[6] Consider the famous (and oft cited) definition of prejudice provided by Allport (1954): "Ethnic prejudice is an antipathy based on a faulty and inflexible generalization" (p. 9). Key to this definition is the idea that pure antipathy and beliefs about other groups are intertwined. As Katz (1991) points out: "[According to Allport,] What separated a prejudice from other negative social attitudes was, first, the inaccuracy of the belief component, which presumably was a consequence overgeneralization from a set of limited observations." (p. 131).

[7] If one group is much larger, then all things equal we would expect the police to spend more time policing that group. The disparities of concern are really with respect to time spent policing per individual. Accounting to this would add complexity to the model without obviously changing our results.

can't choose to allocate *all* of his time to one group or the other. Formally:

Assumption 1. *The officer chooses* $w_A \in [\underline{w}, \overline{w}]$*, where* $0 < \underline{w} \leq 1/2 \leq \overline{w} < 1$.

In addition to being realistic (as we discuss below), both aspects of this assumption—that equal policing is feasible and only policing one group is infeasible—reduce the number of cases to consider for some of our results.[8] We also introduce terminology for the important case where the officer polices one group as much as is feasible:

Definition 1. *If the officer's policing allocation is at either bound* ($w_A \in \{\underline{w}, \overline{w}\}$)*, we say he engages in "extreme policing."*

In the United States, it is typically illegal for governments (including police departments) to target individuals solely on the basis of their social grouping, such as their race, religion, gender, etc. Thus, one way to think about the choice in our model is that the police department decides to target resources toward different geographical locations, which due to residential segregation, have different proportions of the two groups. Unless geographic segregation is absolute, sending officers only to some geographic areas won't completely prevent officers from coming to contact with individuals from both groups. Due to this, the department always has some leeway in determining how much officers come into contact each group, but it cannot choose to allocate all of those officers' time to one neighborhood or another. In Appendix A, we provide a microfoundation for the officer's choice in which the officer chooses how to allocate time between neighborhoods, and not between social groups.

We assume that the allocation of policing effort, as reflected by w_A, affects the detection of crime. As a result, one way to view our model is as a model of "proactive policing," rather than "reactive policing" where officers respond to reports of crimes in progress or which have already

[8] For example, ruling out $w_A = 0$ or $w_A = 1$ removes corner solutions where the officer spends no time policing one of the groups and, as a result, believes that group commits no crime.

occurred (e.g., via 911 calls). Our model is less applicable for crimes that are universally (or near universally) reported, such as murder. More generally, what matters for our argument is that police detect more crime among groups that commit crimes at a higher rate, and where they spend more resources policing.

Formally, we let the amount of crime caught among members of group J be $c_J = p_J w_J$, where $p_J > 0$. The simplest way to interpret this is that p_J represents the average number of crimes committed by members of group J per unit of time, and w_J represents how much time is spent policing this group. This is the data that the officer uses to determine how to allocate his time. In Appendix E, we analyze a variant where the number of crimes caught is not linear in w_J, which complicates the interpretation of the parameters, but does not fundamentally change our argument.

Preferences We assume that objective of the officer is to catch crimes. To capture the notion that the officer might have a taste for discrimination, we allow him to prefer catching crimes among one group or the other. We also assume that there are diminishing returns to the amount of crime caught within each group. This assumption is a reduced-form way to capture the notion that some crimes are "more important" to detect than others, and that the officer will first dedicate time to detecting the more important crimes (within each group).

In Appendix D, we consider more general preferences to capture these ideas, but in the main analysis we use following utility function:

$$u(c_A, c_B) = t_A\sqrt{c_A} + t_B\sqrt{c_B} = t_A\sqrt{p_A w_A} + t_B\sqrt{p_B(1 - w_A)} \tag{1}$$

where $t_A > 0$ and $t_B > 0$ each represent the officer's "taste" for catching crimes among group A and B, respectively. We formally capture the possibility that an officer has animus toward group J by allowing his taste for catch crimes among group J (t_J) to be higher than his taste for catching crimes among the other group.[9]

Definition 2. *The officer has* ***animus towards group*** A *if* $t_A > t_B$*, and* ***animus towards group*** B *if* $t_B > t_A$*. He has no animus if* $t_A = t_B$*.*

We purposefully use the term "animus" instead of the more common term "prejudice" to refer to the component of our model that generates discrimination via preferences. As we discuss above, the use of the term "prejudice" in prior research (and in popular discourse) frequently conflates the preference and belief mechanisms, and we want to keep these separate in our model's primitives. However, our analysis will demonstrate that those with more animus towards a group will also tend to have incorrect beliefs that members of that group commit more crimes.

2.1 Full Information Policing

If the officer knows the true crime rates p_A and p_B (and his own t_A and t_B parameters), maximizing his utility function is straightforward. He may choose to allocate as much time as possible to policing one group or the other (i.e., $w_A = \underline{w}$ or $w_A = \overline{w}$), or an amount that is between these extremes, i.e. an interior allocation. Taking the derivative of u with respect to w_A and setting this equal to zero gives a unique candidate for an interior allocation:

$$w_A^{\dagger}(r_t, r_p) = \frac{r_t^2 r_p}{1 + r_t^2 r_p} \tag{2}$$

where we have reduced notation by defining:

$$r_t = \frac{t_A}{t_B} \quad \text{and} \quad r_p = \frac{p_A}{p_B}$$

[9] In some contexts, and especially outside the U.S., animus toward a group might manifest as a preference for *underpolicing* that group. However, most of the concern about disparities in policing in the U.S. focuses on the ways that some groups (e.g., non-white citizens) are *overpoliced*.

These two parameters correspond exactly to the standard explanations for discrimination. The r_t parameter reflects the relative preferences for catching crimes among each group, which captures the possibility of taste-based discrimination. Given our definition above, an officer with $r_t > 1$ has animus towards group A, and $r_t < 1$ indicates animus towards group B. The r_p parameter reflects the *true* ratio in crime rates of the two groups, capturing the possibility of statistical discrimination. That is, if $r_p > 1$, the crime rate among members of group A is higher than the crime rate among members of group B, and if $r_p < 1$, the crime rate among members of group B is higher than the crime rate among members of group A. When it does not cause confusion, we suppress the r_t and r_p arguments when writing $w_A^\dagger$.

Since $r_t > 0$ and $r_p > 0$, it follows from (2) that $0 < w_A^\dagger < 1$. However, recall that the officer's choice must be between $\underline{w} > 0$ and $\overline{w} < 1$, so it is possible that the utility-maximizing policing choice is outside these bounds, which results in extreme policing, i.e. a "corner solution." In order to reduce the number of cases we need to consider in our analysis, we will assume that the parameters of the model are such that this does not occur when the officer has full information:

Assumption 2. *If the officer has full information, then policing is non-extreme ($\underline{w} < w_A^\dagger < \overline{w}$).*

Given this assumption, an officer with full information will choose allocation $w_A^\dagger$, which we hereafter call the *full information policing*. This policing choice is increasing in r_t, meaning officers who have more animus towards group A will police this group more intensely. It is also increasing in r_p, meaning officers will spend more time policing a group when they believe crime is more prevalent among members of that group. Note that with full information policing, this belief will be based on *actual* crime rates. However, in our main analysis below, this belief may be distorted away from actual crime rates.

While the full information benchmark policing choice is "optimal" given the specified utility function for the officer, we should be clear that it is often not optimal for the policed communities, or society in general. As an extreme example, it could be the case that crime is more prevalent among members of group A, but that the officer has such strong animus towards group B that group B ends up being policed much more heavily.

Whenever the officer polices one group more than the other group, there is a *policing disparity*, given by:

$$\Delta^\dagger \equiv |w_A^\dagger - 1/2|.$$

Since our model allows for both taste-based and statistical discrimination (via parameters r_t and r_p), $\Delta^\dagger$ can be decomposed into two component parts. Formally, define $w_A^{\text{stat}} = w_A^\dagger(1, r_p) = r_p/(1 + r_p)$ to be the "statistical policing" allocation, which reflects what an officer does if he has no animus toward either group but statistically discriminates based on differences in the (true) crime rates. Then, the extent of statistical discrimination is captured by $w_A^{\text{stat}} - 1/2$ and the extent of taste-based discrimination is captured by $w_A^\dagger - w_A^{\text{stat}}$. Taken together, the policing disparity under full information policing can be decomposed as follows:[10]

$$\Delta^\dagger = |\underbrace{(w_A^\dagger - w_A^{\text{stat}})}_{\substack{\text{taste-based}\\\text{discrimination}}} + \underbrace{(w_A^{\text{stat}} - 1/2)}_{\substack{\text{statistical}\\\text{discrimination}}}| = |w_A^\dagger - 1/2|$$

While it should be uncontroversial that taste-based discrimination is normatively undesirable, the normative desirability of statistical discrimination is less clear. Focusing policing efforts toward communities with higher crime rates has the potential to make those communities safer. It may also increase community engagement if citizens do not perceive it to be too invasive (Lerman and Weaver 2014) and increase citizens' willingness to cooperate with police if community-oriented policing tactics are used (Peyton, Sierra-Arévalo, and Rand 2019). On the other hand, there is no

[10] Note that taste-based and statistical discrimination may yield discrimination against different groups. In this case, the policing disparity under full information will be closer to zero than the disparities generated by either kind of discrimination on its own.

guarantee that the "socially optimal" policing allocation corresponds to one in which an officer targets his efforts to higher crime communities. First, statistical discrimination can lead to inefficient stereotyping (see, for example, Coate and Loury 1993; Harcourt 2007; Glaser 2015; O'Flaherty and Sethi 2019). Second, "over-policing" certain social groups can have other spillovers, from reducing political participation (Weaver and Lerman 2010) to reinforcing those groups' status as "race-class subjugated" communities whose primary interaction with the state involves negative interactions with police (for an overview, see Soss and Weaver 2017).

2.2 Behavioral Policing

We now turn to our main analysis, which considers a situation in which the officer does not know the relative crime rates of the two groups (r_p), and forms this belief based on data generated by his policing choices. In reality, police departments collect data on crime from a wide variety of sources. For example, the department's crime statistics may include data about complaints, arrests and possibly even surveys of residents (such as the National Crime Victimization Survey). In this section, we assume that the data the officer in our model uses is entirely driven by the crime detected as a result of his policing choices. In Section 3, we explore the consequences of officers making choices based on data generated by *other* officers' choices as well.

A natural "inferential mistake" the officer might make is to assume that the amount of crime detected among members of each group reflects the crime rates of those groups, without accounting for the fact that one group may be more heavily policed than the other. As this involves the officer forming a posterior belief without conditioning on all relevant information, it is *non-conditioning bias*.

We will allow the officer's non-conditioning bias to be more or less severe. Formally, we use the parameter $\nu \in [0, 1]$ to scale this severity. In the extreme where $\nu = 1$, he may simply compute the ratio of the number crimes detected among members of each group:

$$\tilde{r}_p(w_A, \nu = 1) = \frac{c_A}{c_B} = \frac{p_A w_A}{p_B(1 - w_A)} \tag{3}$$

Note that equation (3) is increasing in w_A, which implies that if the officer forms his belief about r_p in this fashion, he will think that there is a higher crime rate among members of group A (relative to group B) whenever that group is policed more heavily.

The correct way to form a belief about crime given the available data (c_A, c_B, w_A, and w_B) is to divide the crimes caught by the amount of time spent policing each group before making the comparison; i.e., to properly condition. When doing so (and with a large amount of data), the officer will form an accurate belief about the groups' crime rates. Formally, if $\nu = 0$ indicating that does not have any non-conditioning bias, then:

$$\tilde{r}_p(w_A, \nu = 0) = \frac{\frac{p_A w_A}{w_A}}{\frac{p_B(1-w_A)}{1-w_A}} = \frac{p_A}{p_B} = r_p \tag{4}$$

Expressions (3) and (4) respectively correspond to the two extreme situations: the officer makes the worst kind of inferential mistake and the officer does not make this inferential mistake at all. In general, we allow the officer to have a milder form of non-conditioning bias by introducing a parameter $\nu \in [0, 1]$, and the more general belief is:

$$\tilde{r}_p(w_A) = \frac{\frac{p_A w_A}{\nu+(1-\nu)w_A}}{\frac{p_B(1-w_A)}{\nu+(1-\nu)(1-w_A)}} = r_p \left(\frac{w_A(\nu + (1 - \nu)(1 - w_A))}{(1 - w_A)(\nu + (1 - \nu)w_A)} \right). \tag{5}$$

For conciseness, we will suppress the ν argument of $\tilde{r}_p$ in the remainder of the analysis. As ν approaches zero, the officer's belief about crime, $\tilde{r}_p$, becomes more accurate (i.e., approaches r_p). As ν approaches one, $\tilde{r}_p$ approaches the belief formed by the most extreme non-conditioning bias. More generally, as ν increases, the officer makes a more severe inferential mistake.

That police officers exhibit non-conditioning bias is not an outlandish proposition. Consider several examples. Using a case study of drug arrests in Oakland, California, Lum and Isaac (2016) demonstrate that data used in predictive policing algorithms perpetuates policing disparities since it is based on past policing patterns and does not appear to reflect *actual* drug use patterns. In her opinion in *Floyd v. New York*, U.S. District Judge Scheindlin writes "The City and its highest officials believe that blacks and Hispanics should be stopped at the same rate as their proportion of the local criminal suspect population" (p. 9). This is a prime example of non-conditioning bias, which is precisely what Judge Scheindlin finds troubling: "Instead, I conclude that the benchmark used by plaintiffs' expert—a combination of local population demographics and local crime rates (*to account for police deployment*) is the most sensible" (p. 9, emphasis added). Finally, Glaser (2015) recounts a particularly clear example of non-conditioning bias when a former Los Angeles police chief told a reporter: "if officers are looking for criminal activity, they're going to look at the kind of people who are listed on crime reports" (p. 96). Of course, the "kinds of people who are listed on crime reports" will be disproportionately from highly policed communities and not necessarily representative of those who are prone to commit crimes.[11]

Equilibrium Thus far, we have described how the officer's belief depends on his policing decisions, given that he may have a non-conditioning bias that causes him to make an inferential mistake. We next consider how the officer's policing decisions depend on this potentially inaccurate belief. If the officer's policing decision and his (potentially inaccurate) belief are mutually reinforcing, then we will call this an "equilibrium" of the model since it serves as a useful prediction about what the officer would do.

[11] While examples of non-conditioning bias abound, we make no specific claim about the severity of this bias across contexts. Individuals vary with their ability to make accurate inferences from data, and police departments use a variety of statistics, some of which may not be affected by non-conditioning bias. For this reason, we allow for relatively mild or severe forms of the bias, as represented by $\nu \in [0, 1]$.

Formally, the officer will choose the policing allocation w_A that is obtained by maximizing his utility given his belief $\tilde{r}_p(w_A)$ as defined in (5). This is his "best response" to his beliefs, which is given by:

$$w_A^{\text{br}}(r_t, \tilde{r}_p) = \begin{cases} \underline{w} & \text{if } \frac{r_t^2 \tilde{r}_p}{1+r_t^2 \tilde{r}_p} < \underline{w} \\ \frac{r_t^2 \tilde{r}_p}{1+r_t^2 \tilde{r}_p} & \text{if } \frac{r_t^2 \tilde{r}_p}{1+r_t^2 \tilde{r}_p} \in [\underline{w}, \overline{w}] \\ \overline{w} & \text{if } \frac{r_t^2 \tilde{r}_p}{1+r_t^2 \tilde{r}_p} > \overline{w} \end{cases} \tag{6}$$

Note that this best response resembles the full information case, but with the potentially incorrect belief $\tilde{r}_p$ replacing the true ratio r_p, and with the possibility of extreme policing (i.e., $\underline{w}$ or $\overline{w}$).

We have characterized how his beliefs respond to his actions and how his actions respond to his beliefs. We now formally define what constitutes an equilibrium to the model we analyze in this section.

Definition 3. *An* ***equilibrium of the unitary officer model*** *is a policing allocation w_A^* and a belief about crime rates $\tilde{r}_p^*$, where*

(i) w_A^* solves $w_A^* = w_A^{br}(r_t, \tilde{r}_p^*)$; and

(ii) $\tilde{r}_p^* = \tilde{r}_p(w_A^*)$.

An equilibrium of the model exists and is unique. Accordingly, one of our main innovations is to demonstrate that it is possible to obtain an equilibrium of a model of policing in which an officer forms non-standard (and potentially inaccurate) beliefs. Put another way, even though more intense policing of one group creates a feedback channel, it is not the case that *any* policing allocation can be self-enforcing. To see why, note that there is a unique solution to $w_A = \frac{r_t^2 \tilde{r}_p(w_A)}{1+r_t^2 \tilde{r}_p(w_A)}$ given by:

$$\widehat{w}_A = w_A^{\dagger} + \frac{\nu(r_t^2 r_p - 1)}{(1-\nu)(1+r_t^2 r_p)} \tag{7}$$

If $\widehat{w}_A$ lies in $[\underline{w}, \overline{w}]$, then it corresponds to an equilibrium allocation. Whenever $\widehat{w}_A$ does not lies in $[\underline{w}, \overline{w}]$, then there is an equilibrium involving extreme policing (i.e., a corner solution):

Proposition 1. *There is a unique equilibrium in which the officer chooses a policing allocation*

$$w_A^* = \begin{cases} \underline{w} & \textit{if } \widehat{w}_A < \underline{w} \\ \widehat{w}_A & \textit{if } \widehat{w}_A \in [\underline{w}, \overline{w}] \\ \overline{w} & \textit{if } \widehat{w}_A > \overline{w} \end{cases}$$

and forms a (potentially inaccurate) belief $\tilde{r}_p^*$ using (5).

Proof All proofs are in the appendix.

A natural way to conceptualize an equilibrium is in a dynamic setting. An officer chooses a policing allocation for some "time period," and then forms an updated belief about crime rates using (5) and given the data generated by his policing allocation. If the officer wants to change his policing allocation given this updated belief— i.e., $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A)) \neq w_A$—then he is not in an equilibrium. If the officer wants to continue to use the same policing allocation given his updated belief—i.e., $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A)) = w_A$—then he is in an equilibrium.

Figure 1 illustrates this dynamic process. In each panel (which vary in their values of r_t and r_p), the black curves trace out $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A))$ as a function of w_A. The grey 45-degree line represents points where the best response allocation equals the actual allocation. Starting at any point w_A, if the $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A))$ curve lies above the 45 degree line, then an officer who initially policies at allocation w_A will generate a belief about the relative crime rates that makes him want to police group A more. Conversely, if the curve lies below the 45 degree line, an officer starting at w_A would want to police group A less. An equilibrium occurs at an intersection of the black curve and the 45 degree line, where the officer would not want to change his allocation.

Setting aside the rest of the markings on the graph for the moment, note that in all but the bottom right panel the intersection is an interior allocation, which identifies a unique equilibrium.

Before we explore how the officer's choice generates discrimination, we first note an important property of the officer's equilibrium policing allocation. As we show in Appendix D.2, the equilibrium allocation in Proposition 1 is "stable" in the sense that it is not sensitive to small perturbations.[12] Visually, this is because the best response curve lies below the 45 degree line above the equilibrium, and above the 45 degree line before the equilibrium. Roughly speaking, in the context of our model, this means that if the officer "accidentally" were to police one group a little more (or a little less) than their equilibrium allocation prescribes, the best response to his new belief (induced by the mistake) would be to move back towards the equilibrium.

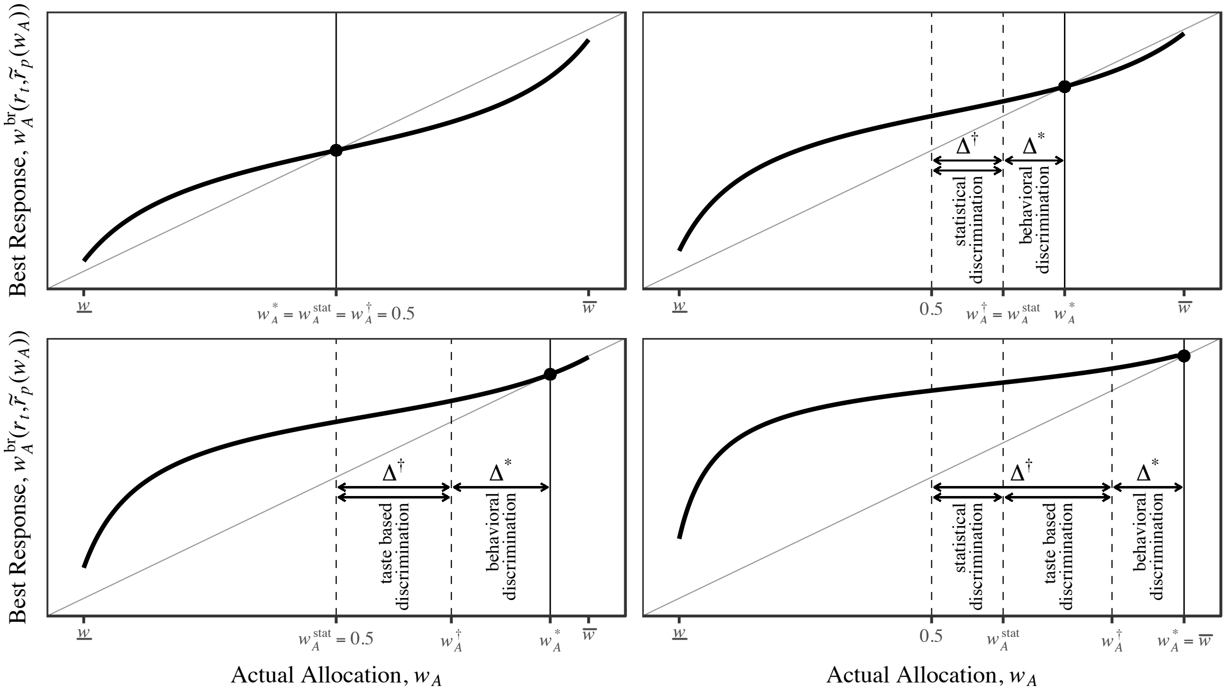
The difference between the panels in figure Figure 1 is that in the top panels the officer has no animus towards either group ($r_t = 1$), while in the bottom panels he has animus towards group A. In the left panels there is no difference in actual crime rates ($r_p = 1$), while in the right panels the true crime rate is higher in group A ($r_p > 1$).

Combining, the top left panel depicts a scenario with equal crime rates and no officer animus, $r_p = r_t = 1$. In this situation, despite making inferential mistakes, the officer's policing allocation is equal, $w_A^* = 1/2$. If the officer were to police group A more or less, there would be "self-correction" in the sense described above: he would move back towards the equilibrium with equal policing.

However, equal policing is fragile to changes in the exogenous parameters r_t and r_p. The bottom left panel demonstrates a situation with equal crime rates, but where the officer has animus toward group A. As the figure depicts, without making an inferential mistake, the officer's animus toward group A causes him to engage in taste-based discrimination against group A so that $w_A^\dagger >$

[12] In many models of sorting and statistical discrimination, there are multiple equilibria, some of which are unstable (see, for example, Coate and Loury 1993; Benabou 1993; Chaudhuri and Sethi 2008). Unstable equilibria are undesirable because they only exist if the parameters of the model are exactly right. In the real world, people sometimes make small errors when making decisions, and so it is useful to know that an equilibrium will persist even when these small mistakes occur.

Figure 1: In each panel, we plot the officer's best response $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A))$ as a function of his actual policing allocation w_A. an equilibrium of the model occurs where $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A))$ intersects the diagonal line—i.e., at a fixed point, denoted by a large dot. Each panel depicts equilibria for different parameter values. We also depict the disparities caused by statistical, taste-based and behavioral discrimination in each equilibrium. For the left panels, crime rates are equal ($r_p = 1$) and for the right panels, group A's crime rate is higher ($r_p > 1$). For the top panels, the officer has no animus ($r_t = 1$) and for the bottom panels, the officer has animus against A ($r_t > 1$).



$1/2$ (and thus $\Delta^\dagger > 0$). However, his non-conditioning bias causes him to police group A even more than he would due to his animus alone, $w_A^* > w_A^\dagger$.

Formally, if the officer chooses a policing allocation w_A^* in an equilibrium, then we define the policing disparity relative to full information policing as:

$$\Delta^* \equiv |w_A^* - w_A^\dagger|.$$

This is the "excess disparity" caused by the fact that the officer makes an inferential mistake when forming his belief about the two crime rates. As a result, we refer to it as *behavioral discrimination*.

We will show below that behavioral discrimination always goes "in the same direction" as the disparity caused by the standard explanations (and represented by $\Delta^{\dagger}$). We can therefore denote total discrimination as $\Delta = \Delta^{\dagger} + \Delta^{*}$. Returning to the bottom left panel of figure Figure 1, in this equilibrium about half of the officer's discrimination is driven by taste and about half is driven by non-conditioning bias.

Behavioral discrimination can also occur in the absence of officer animus. The top right panel indicates a case where $r_t = 1$ but $r_p > 1$. So, some excess policing of group A is explained by different crime rates (again $w_A^{\dagger} > 1/2$, and $\Delta^{\dagger} > 0$), but the officer believes these differences are bigger than they really are. As with the illustration of taste-based discrimination, this force roughly doubles the policing disparity relative to full information policing.

Finally, the bottom right panel shows a case where group A has a higher crime rate and the officer has animus towards this group. In this case, no matter what feasible allocation he chooses, he would always like to police group A even more. This leads to extreme policing of this group even though his policing allocation would be interior if he had full information. As demonstrated below, such extreme policing does not require both officer animus and differential crime rates, but will generically occur as long as the officer's non-conditioning bias is sufficiently strong.

The officer's non-conditioning bias creates a link between taste-based and statistical discrimination. For an officer with any strictly positive level of this bias, taste-based and statistical discrimination are no longer two mutually exclusive channels through which policing disparities emerge. In fact, what we term "behavioral discrimination" is formally equivalent to the excess statistical discrimination that is caused by exaggerated beliefs about relative crime rates. When conceptualized in this way, our model shows that taste-based discrimination can *cause* a (inaccurate) statistical discrimination. And since an officer's animus can cause distorted beliefs about crime rates, our model maps into an intuition in the academic literature (and in popular discourse) that the empirical phenomenon of prejudice will typically involve both racial animus and incorrect beliefs.

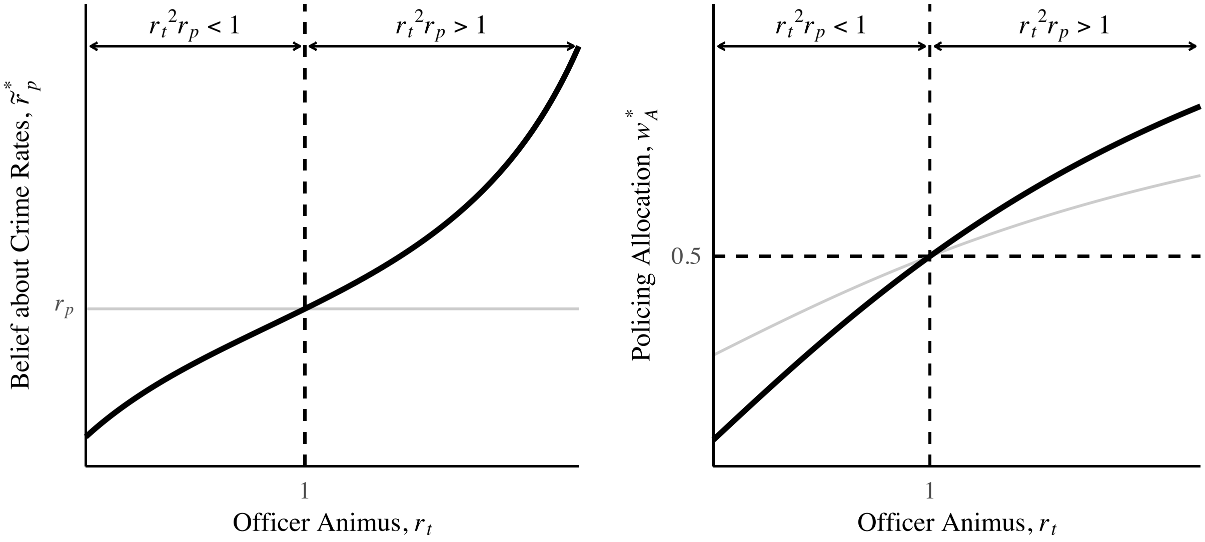
To be more concrete about how this works in our model, consider the following. First, the officer's animus causes him to allocate more policing effort toward one group. Then, since he spends more time policing that group, he sees more crimes among members of that group. Finally, his non-conditioning bias causes him to infer that the increased number of crimes he observes is an indication that the crime rate among members of that group is higher than it actually is. As a result (and notwithstanding his animus), his non-conditioning bias causes him to to *sincerely believe* that some (or even most) his overpolicing of one group is justified by the prevalence of crime among members of that group. As Gelman, Fagan, and Kiss (2007) point out: "Police often point to the high rates of seizures of contraband, weapons, and fugitives in such stops, and also to a reduction of crime, to justify such aggressive policing" (p. 814).

More generally, as the officer's animus increases, then his non-conditioning bias causes increasingly distorted beliefs and increasing levels of behavioral discrimination. In Figure 2, we plot both his equilibrium belief and his equilibrium policing allocation as a function of his animus. The solid grey lines indicates his equilibrium belief and allocation under full information policing (i.e., if he did not have a non-conditioning bias), and the solid black lines indicates his equilibrium belief and allocation when he has a non-conditioning bias.

Note that without non-conditioning bias, his belief does not depend on his animus and remains constant (at the true crime rate) no matter how much animus he has. In other words, taste-based and statistical discrimination are independent of one another with full information policing. However, once he has a non-conditioning bias, as r_t increases he forms increasingly exaggerated beliefs about the relative crime rate among members of group A. This causes his policing allocation to be even more unequal than it would under full information policing, as the right panel shows.

While we have depicted how non-conditioning bias amplifies taste-based discrimination, it is also the case that it amplifies discrimination due solely to differences in crime rates. In other words, if crime rates are not equal between groups, non-conditioning bias causes even statistical discriminators (i.e., those with no animus) to overpolice one group as though they had animus toward that group.

Figure 2: The solid grey lines depicts his policing allocation and belief about relative crime rates under full information policing, and the solid black lines depicts his policing allocation and belief about relative crime rates under behavioral policing. As the officer's animus toward group A increases, he polices group A more and forms an increasingly exaggerated (and incorrect) belief that crime is relatively more prevalent among members of group A.



The previous analysis suggests that behavioral discrimination will tend to amplify policing disparities caused by taste-based and/or statistical discrimination. We now explore exactly when and how this amplification occurs.

There is one situation where behavioral policing does not amplify policing disparities.[13] If there would be no policing disparity with full information policing, then behavioral policing cannot

[13] If we relax Assumption 2, then there is a second situation in which there is no behavioral discrimination. If the officer would engage in extreme policing regardless whether he has a non-conditioning bias, then trivially, behavioral policing does not increase the policing disparity. Formally, this occurs if $w_A^* = w_A^\dagger = \overline{w}$ or $w_A^* = w_A^\dagger = \underline{w}$. However, this scenario is qualitatively less interesting since it occurs solely due to the fact that the officer's choice is constrained to be in $[\underline{w}, \overline{w}]$ and there is no "room" for him to discriminate any more than he otherwise would with full information.

amplify policing disparities. Formally, this occurs if $r_t^2 r_p = 1$ so that $w_A^* = w_A^\dagger = 1/2$. In this case, behavioral policing does not *by itself* lead to discrimination against one group. One way that this may occur is if the officer has no animus towards either group ($r_t = 1$) and crime is group-invariant ($r_p = 1$). It is also possible if the officer has animus towards one group but the other group has a crime rate just high enough to exactly offset the officer's animus—formally, this occurs if $r_p = 1/r_t^2$.

However, this is a very specific situation. If $r_t^2 r_p < 1$, then $w_A^* < w_A^\dagger$, meaning the officer polices group A less than he would with full information (and polices group B more). Conversely, when $r_t^2 r_p > 1$, the officer polices group A more than he would with full information (and polices group B less), $w_A^* > w_A^\dagger$. In both cases, $\Delta^* > 0$, meaning that behavioral policing generically amplifies whatever disparities would exist if officer formed accurate beliefs about crime rates.

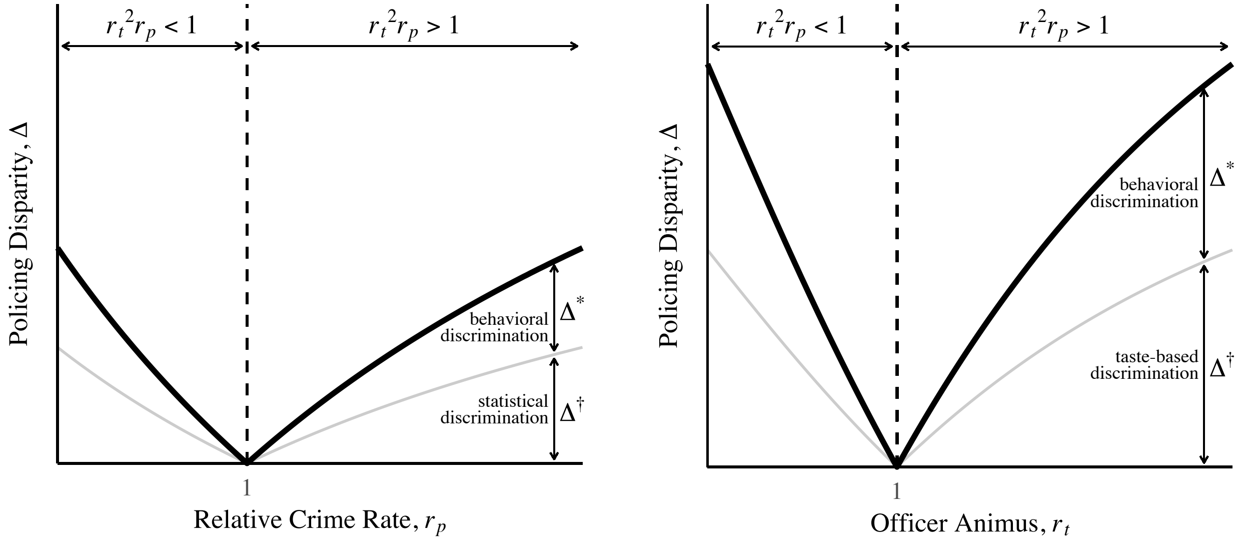
Proposition 2. *For any $\nu \in (0,1)$:*

(i) If $r_t^2 r_p = 1$, then there is no policing disparity (since $w_A^* = w_A^\dagger = 1/2$), and the officer has correct beliefs about crime, $\tilde{r}_p^* = r_p$.

(ii) If $r_t^2 r_p \neq 1$, then behavioral policing amplifies existing disparities: $w_A^* > w_A^\dagger > 1/2$ if $r_t^2 r_p > 1$ and $w_A^* < w_A^\dagger < 1/2$ if $r_t^2 r_p < 1$ (alternatively, $\Delta^* > 0$), and the officer has incorrect beliefs, $\tilde{r}_p^* \neq r_p$.

If the officer's policing allocation is not extreme, then the disparity caused by behavioral discrimination is strictly positive as as $r_t^2 r_p$ moves away from 1. Figure 3 illustrates. In the left panel, we plot policing disparities as a function of the (true) relative crime rate, r_p. In the right panel, we plot policing disparities as a function of the officer's animus, r_t. In each panel, the grey line depicts the policing disparity caused by statistical and taste-based discrimination and the black line depicts the entire policing disparity. Note that in either panel, as long as $r_t^2 r_p \neq 1$, then behavioral discrimination causes the policing disparity to be higher than it otherwise would have been with only taste-based and statistical discrimination.

Figure 3: In each panel, we plot the policing disparity that emerges in a non-extreme equilibrium of the model, as a function of the true relative crime rate (left panel) and the officer's animus toward group A (right panel). As long as $r_t^2 r_p \neq 1$, the officer always engages in either statistical or taste-based discrimination, *as well as* behavioral discrimination.



For Proposition 2 and Figure 3, we assume that ν is intermediate. As we demonstrate in the next result, as the severity of the officer's inferential mistake (as reflected by the value of ν) increases, the policing disparity increases until the officer reaches extreme policing. Strikingly, as long as behavioral policing causes any *any* disparity, then as the officer's inferential mistake become more severe, it eventually leads him to engage in extreme policing.

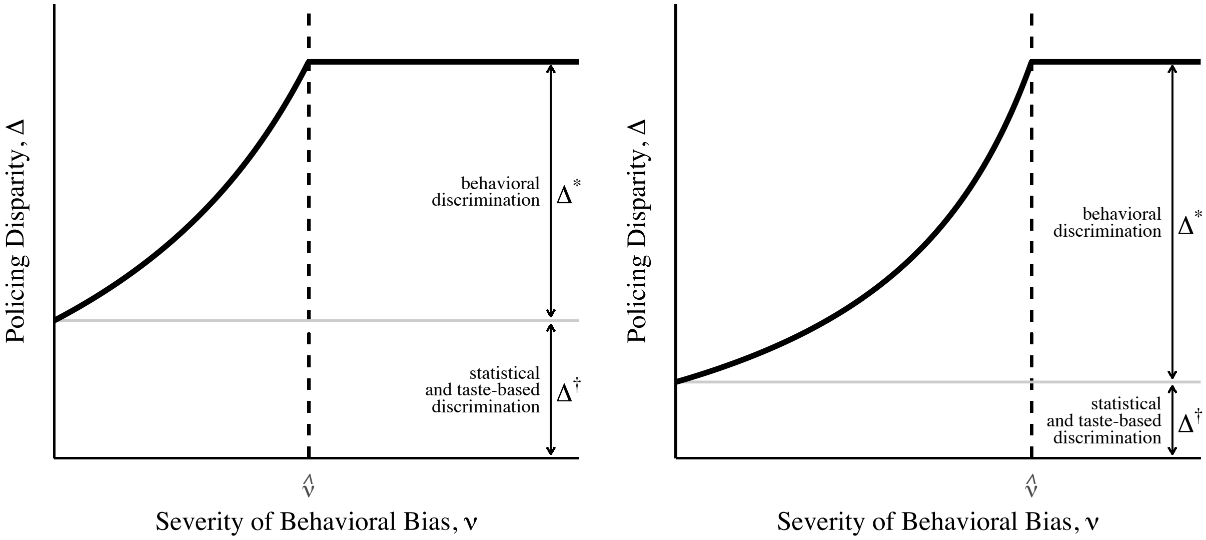
Proposition 3. *Assume $r_t^2 r_p \neq 1$. Then:*

(i) When the officer does not engage in extreme policing ($\underline{w} < w_A^* < \overline{w}$), the policing disparity caused by behavioral discrimination (Δ^*) is strictly increasing in ν.

(ii) There exists a $\widehat{\nu} < 1$ such that if $\nu \geq \widehat{\nu}$, the officer engages in extreme policing of one group and the policing disparity is at its maximum, i.e. $w_A^* = \underline{w}$ or $w_A^* = \overline{w}$.

Figure 4 depicts the result for two situations. In each panel, the x-axis plots ν and the y-axis plots the policing disparity. The left panel depicts a situation where statistical and taste-based discrimination cause a large disparity, and the right panel depicts a situation where statistical

Figure 4: As the severity of the officer's inferential mistake (i.e., ν) increases, so do policing disparities caused by behavioral discrimination. Moreover, very severe non-conditioning biases can cause such severe behavioral discrimination that the officer engages in extreme policing when he would not do so under full information policing.



and taste-based discrimination cause a small disparity. In either case, at any interior allocation behavioral discrimination strictly increases as ν increases, and if ν is sufficiently high the officer will engage in extreme policing.

In this section, we have analyzed a model of policing by a representative officer. With our analysis, we make three main points. First, we demonstrate that there is a policing allocation in which an officer makes an inferential mistake that affects their policing, and where their policing decision feeds back into and reinforces their (incorrect) beliefs. If the officer makes an inferential mistake when making decisions about how to allocate policing resources, we call this *behavioral policing*.

Second, behavioral policing generically causes the officer to engage in behavioral discrimination in which he overpolices one group because he forms an exaggerated belief that the relative crime rate among members of that group is higher than it actually is. An implication of this is that taste-based discrimination can perversely cause (inaccurate) statistical discrimination. Recall we previously noted that use of the term "prejudice" in prior work and popular discourse often conflates preferences and beliefs. In a sense, our results provide a justification for conflating these two mechanisms, as they will inevitably go together for those who make the kind of inferential mistake we study. However, the link between preferences and beliefs requires people to process information incorrectly. So, if reducing animus itself is not possible, it may still be possible to reduce prejudice by training officers to engage in more accurate statistical reasoning about why they may be observing cross-group differences in crime (or other relevant data).

Third, we showed that behavioral discrimination amplifies the policing disparity caused by taste-based and/or statistical discrimination. In particular, if $\nu > 0$ and $r_t^2 r_p \neq 1$, then he will always overpolice one group more than he otherwise would with full information. Moreover, the extent to which behavioral discrimination amplifies existing a policing disparity increases with the severity of his inferential mistake. As a result, behavioral policing can sometimes dramatically increase policing disparities, even inducing the officer to engage in extreme policing of one group.

Our model provides new insights about the causes of policing disparities. However, one potential limitation is that the data collected from the officer's own policing decisions (i.e. w_A^*) is the only thing causing him to form distorted beliefs. In reality, police departments are comprised of a multiple police officers with diverse preferences, and all of their individual policing choices end up contributing to the department's overall assessment of crime across communities. In the next section, we extend the model to look at how the presence of multiple, heterogeneous police officers affects our findings. Strikingly, we show that one officer's animus can "spill over" to a second officer, distorting that second officer's belief about crime rates and causing them to discriminate.

3 Model with Multiple, Heterogeneous Officers

To study how the dynamics of the model are different with multiple decision-makers, we analyze the simplest such environment: with two officers, indexed by $i \in \{1, 2\}$. Both officers choose how much time to allocate to group A, $w_{A,i} \in [\underline{w}, \overline{w}]$, with the remainder allocated to group B: $w_{B,i} = 1 - w_{A,i}$. In this section, let $w_J = w_{J,1} + w_{J,2}$ represent the *total* policing of group J. (Note that in this section, $2\underline{w} \leq w_J \leq 2\overline{w}$, since there are two officers each allocating 1 unit of time.) Let $c_{J,i} = p_J w_{J,i}$ be the number of crimes caught among group J by officer i, and $c_J = p_J w_J$ the total crime caught among members of group J.

We assume each officer cares only about the number of crimes that he catches:

$$u_i(c_{A,i}, c_{B,i}) = t_{A,i}\sqrt{c_{A,i}} + t_{B,i}\sqrt{c_{B,i}} = t_{A,i}\sqrt{p_A w_{A,i}} + t_{B,i}\sqrt{p_B(1 - w_{A,i})} \tag{8}$$

This utility function allows us to isolate the affect of distorted beliefs on policing since it means that there is no *direct* effect of officer j's behavior on the utility of officer i. There will only be an *indirect* effect of the other officer's behavior via officer i's belief. If instead each officer's utility were to be defined over the total crime caught, then the policing behavior of the other officer has a direct effect on his own best response, and we would not be able to cleanly isolate how much distorted beliefs affect policing decisions.

By an identical analysis to the case of the single officer with full information about the crime rates and assuming a non-extreme policing, the best response for each officer i depends on his animus ($r_{t,i}$) and the true ratio of crime rates (r_p):

$$w^{\dagger}_{A,i} = \frac{r^2_{t,i} r_p}{1 + r^2_{t,i} r_p}. \tag{9}$$

Because each officer's utility only depends on the crimes he catches, this allocation does not depend on the beliefs or behavior of the other officer in any way.

We also define the officers' beliefs in a similar way to the single officer model, but accounting for the fact that there are now two officers making policing allocations:

$$\tilde{r}_{p,i}(w_A) = \frac{\frac{c_A}{\nu_i+(1-\nu_i)w_{A,1}+\nu_i+(1-\nu_i)w_{A,2}}}{\frac{c_B}{\nu_i+(1-\nu_i)w_{B,1}+\nu_i+(1-\nu_i)w_{B,2}}} = \frac{\frac{c_A}{2\nu_i+(1-\nu_i)w_A}}{\frac{c_B}{2\nu_i+(1-\nu_i)(2-w_A)}} \tag{10}$$

Note that each officer's belief in the multiple officer model is indexed by i since each officer can, in principle, differ with respect to the severity of their non-conditioning bias (i.e., have different values of ν_i).

This definition implicitly assumes that each officer's failure to correct for policing intensity is symmetric in the sense that they fail to adjust for both their choice and the other officer's choice. In Appendix D.3, we consider an alternative version of this bias where the officers adjust differently for their own behavior and the other officer's behavior. The key property of the symmetric version we study here, as well as the version we study in the appendix, is that officer i's belief about the relative prevalence of crime among members of group A is increasing in how much the *other* officer polices group A.

Each officer's best response also resembles the single officer case:

$$w_A^{\text{br}}(r_{t,i}, \tilde{r}_{p,i}) = \begin{cases} \underline{w} & \text{if } \frac{r_{t,i}^2 \tilde{r}_{p,i}}{1+r_{t,i}^2 \tilde{r}_{p,i}} < \underline{w} \\ \frac{r_{t,i}^2 \tilde{r}_{p,i}}{1+r_{t,i}^2 \tilde{r}_{p,i}} & \text{if } \frac{r_{t,i}^2 \tilde{r}_{p,i}}{1+r_{t,i}^2 \tilde{r}_{p,i}} \in [\underline{w}, \overline{w}] \\ \overline{w} & \text{if } \frac{r_{t,i}^2 \tilde{r}_{p,i}}{1+r_{t,i}^2 \tilde{r}_{p,i}} > \overline{w} \end{cases}$$

which is increasing in both his animus towards group A and his belief about the relative prevalence of crime among members of group A. This observation, combined with the fact that each officer's belief is affected by the policing allocation of the other officer, gives the intuition for the main result in this section. First, as officer 1's animus toward group A increases, this leads him to police group A more heavily (and, as in the previous section, this effect is amplified by inaccurate belief formation). And second, as long as officer 2 does not account for officer 1's animus-driven increased policing of group A, it will also lead officer 2 to believe (inaccurately) that group A has a higher crime rate. As a result, the animus of one officer ends up spilling over into the behavior of the other officer.

In the remainder of this section, we first demonstrate that this property holds in an equilibrium of the model, and then explore some more subtle properties of the resulting policing allocations and beliefs about crime. Formally, we define a solution of the multiple officer model as follows:

Definition 4. *An* ***equilibrium of the model with two officers*** *is a pair of allocation choices* $(w^*_{A,1}, w^*_{A,2})$ *and vector of beliefs* $(\tilde{r}^*_{p,1}, \tilde{r}^*_{p,2})$ *such that for all* $i \in \{1, 2\}$*:*

(i) $w^*_{A,i} = w^{br}_A(r_{t,i}, \tilde{r}^*_{p,i})$*, and*

(ii) $\tilde{r}^*_{p,i} = \tilde{r}_{p,i}(w^*_A)$ *is given by equation (17) evaluated at* w^*_A*.*

(The definition naturally extends to more than two officers.)

With multiple officers, it is difficult to obtain closed-form solutions. However, it is straightforward to show that an equilibrium exists, and in any equilibrium that meets a stability condition analogous to the single-officer model (see Appendix B.2), the comparative statics are consistent with the conjectures above. We are primarily interested in the role that distorted beliefs play in policing disparities. More specifically, we examine how discrimination can "spill over" from one officer to another.

Our main result in the multiple officer model illustrates how this occurs:

Proposition 4. *In the model with two officers, an equilibrium exists. At any stable interior equilibrium allocation:*

(i) Each officer's allocation to group A $(w^*_{A,i})$ *is strictly increasing in the animus of either officer,* $r_{t,1}$ *or* $r_{t,2}$*, and*

(ii) If the officers collectively spend more than half of their time policing group J $(w^*_J > 1)$*, then each officer's allocation to group* J *is strictly increasing in the non-conditioning bias*

of either officer, ν_1 or ν_2.

In words, part (i) states that as either officer has more animus towards a group, *both* officers end up policing that group more. This is because the other officer (whose animus remains unchanged) does not fully correct for how his peer's increased policing of the group inflates the number of crimes caught among members of that group. In this sense, taste-based discrimination is contagious across officers.

Part (ii) states that whenever the officers collectively spend more time policing one group than the other, increasing the non-conditioning bias of either officer makes both officers decide to police that group even more. The intuition comes from the fact that whenever one group is policed more than the other, increasing one officer's non-conditoning bias has a direct effect on how much he polices that group (increasing it), and then spills over into the other officer's behavior. In this sense, inferential mistakes are contagious across officers.

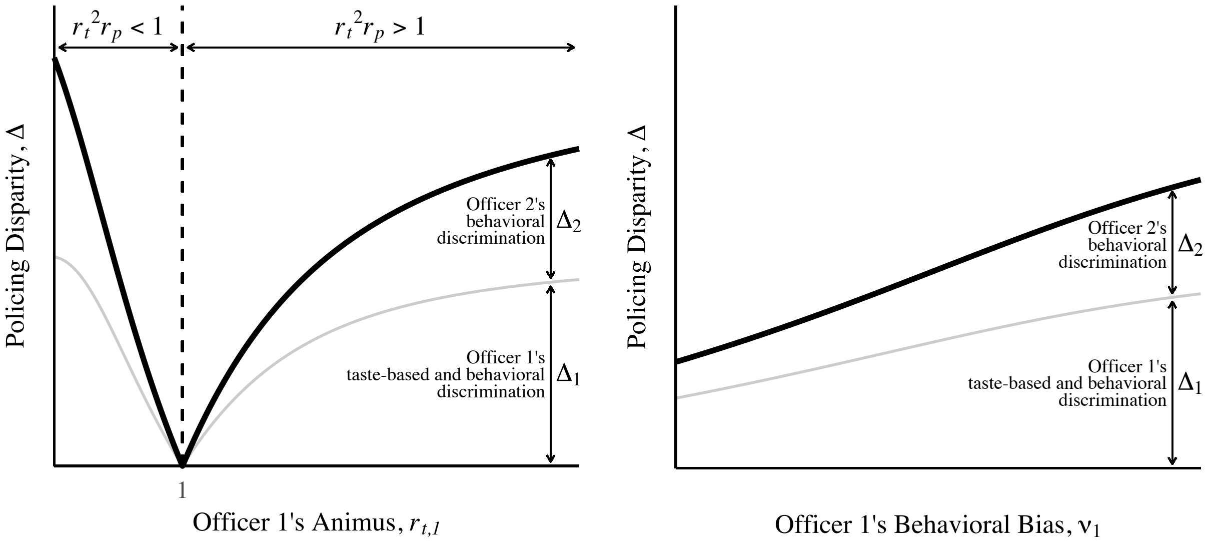
Finally, we illustrate how this affects individual and aggregate discrimination. To consider each officer's discrimination separately, define the policing disparities generated by each officer as follows:

$$\Delta_i^{\dagger} = |w_{A,i}^{\dagger} - 1/2| \qquad\qquad \Delta_i^{*} = |w_{A,i}^{*} - w_{A,i}^{\dagger}|$$

We also define the officer-level discrimination as $\Delta_i = \Delta_i^{\dagger} + \Delta_i^{*}$, and total discrimination by both officers as $\Delta = \Delta_1 + \Delta_2$.

In Figure 5, we depict examples of contagious prejudice (left panel) and contagious inferential mistakes (right panel). For both panels, we assume that crime rates are equal ($r_p = 1$), and that officer 2 has no animus toward either group (so that $r_{t,2} = 1$). The grey dashed line indicates officer 1's policing disparity, which is caused by both taste-based and behavioral discrimination. The black line depicts the total disparity that arises from both officers' policing allocations. The gap between the grey dashed line and the black line gives the disparity caused by officer 2's behavioral discrimination.

Figure 5: When there are multiple officers, discrimination due to animus and inferential mistakes are contagious. In the left panel, we depict how officer 2 discriminates more as officer 1 has more animus. In the right panel, we depict how officer 2 discriminates more as officer 1's non-conditioning bias becomes more severe.



First consider the left plot, which illustrates thesee quantities as a function of $r_{t,1}$. Since officer 2 has no animus and crime rates are equal, he would not discriminate under full information policing even if officer 1 does. However, since officer 2 has a non-conditioning bias, he ends up discriminating *because of* officer 1's animus toward one of the groups. Moreover, the more animus that officer 1 has, the more officer 2 discriminates. Next, consider the right plot, where the x-axis represents ν_1. As officer 1 makes increasingly severe inferential mistakes (holding fixed the officer 2's non-conditioning bias at $\nu_2 = 1/2$), he polices group A more, as he is not accounting for the fact that the higher crime rate among this group is driven by his own animus. This *also* leads officer 2 to police group A more, since he has a non-conditioning bias, $\nu_2 > 0$. In the examples illustrated by both panels, officer 1's discrimination is contagious.

These findings suggest that efforts to reduce policing disparities by reducing officer animus (via training), or diversifying police forces to reduce the number of officers with animus, may be of limited effectiveness as long as some officers still have animus toward one or more groups. Given their non-conditioning bias, a bad apple (or even a well-intentioned, but naive apple) can both spoil the bunch.

4 Conclusion

In this paper, we have provided a unified behavioral theory of discrimination in policing. Our theory is *unified* because it allows for both group-based animus and statistical differences between groups. It is *behavioral* because it relaxes the standard (and unrealistic) assumption that decision makers must be fully Bayesian. In particular, we assume that police officials form beliefs about the relative prevalence of crime among members of two groups without fully accounting for the intensity with which they police each of those two groups. We call this failure to account for policing intensity *non-conditioning bias*.

We show that an officer with this kind of non-conditioning bias will generically overpolice one of two groups due to the fact that he forms exaggerated beliefs about the relative crime rate among members of that group. We call this *behavioral discrimination*. This kind of discrimination will amplify existing disparities caused by taste-based and/or statistical discrimination. Moreover, when an officer has a non-conditioning bias, then it no longer makes sense to treat taste-based and statistical discrimination as separate and independent channels through by which discrimination occurs. Indeed, behavioral discrimination is a form of *inaccurate* statistical discrimination. Our model thus shows how racial animus and discrimination based on incorrect beliefs are intertwined.

We also extend the model to examine how the behavioral policing by multiple officers can generate discrimination. Due to their non-conditioning biases, the group-based animus of one officer can “spill over” and affect the policing decisions of another officer who has no animus toward either group. The analysis suggests that even if a very small number of officers harbor animus and discriminate against one group, other officers may discriminate against that group too.

The mechanism by which our model produces discrimination also potentially sheds light on the source of political and social conflict over biased policing. Many police officials and policing advocates vehemently assert that policing disparities are justified (for many examples, see Gelman, Fagan, and Kiss 2007), while activists and community leaders protest practices they view as racially discriminatory. Our model focuses on the incorrect beliefs formed by police officers, but we should emphasize that all humans can make similar kinds of inferential errors, and these errors can magnify the differences in beliefs for those with different life experiences. More optimistically, if those with different views of what drives discrimination can at least learn to understand where others are coming from, there may be hope for finding mutually acceptable ways to improve policing practices.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. "Explaining Preferences from Behavior: A Cognitive Dissonance Approach." *Journal of Politics* 80 (2): 400–411.

Allport, Gordon W. 1954. *The Nature of Prejudice.* Reading, MA: Addison-Wesley.

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

Arrow, Kenneth J. 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets,* edited by Orley Ashenfelter and Albert Rees. Princeton University Press.

Ash, Elliott, Jeffrey Fagan, and Allison P. Harris. Forthcoming. "Local Public Finance and Discriminatory Policing: Evidence from Traffic Stops in Missouri." *Journal of Race, Ethnicity, and Politics.* https://www.allisonpharris.com/uploads/1/0/7/3/107342067/ash-fagan-harris_6-2-18.pdf.

Balko, Radley. 2018. "There's Overwhelming Evidence that the Criminal-Justice System Is Racist. Here's the Proof." *Washington Post.* https://www.washingtonpost.com/news/opinions/wp/2018/09/18/theres-overwhelming-evidence-that-the-criminal-justice-system-is-racist-heres-the-proof/.

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

Benabou, Roland. 1993. "Workings of a City: Location, Education, and Production." *The Quarterly Journal of Economics* 108 (3): 619–652.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination," NBER Working Paper Series, no. 25935 (June). doi:10.3386/w25935. http://www.nber.org/papers/w25935.

Broockman, David E., and Evan J. Soltas. 2018. "A Natural Experiment on Discrimination in Elections." Manuscript. `https://ssrn.com/abstract=2919664`.

Butler, Daniel M., and David E. Broockman. 2011. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55 (3): 463–477.

Camerer, Colin. 1995. "Individual Decision Making." In *Handbook of Experimental Economics,* edited by John H. Kagel and Alvin E. Roth, 587–703. Princeton University Press.

Chaudhuri, Shubham, and Rajiv Sethi. 2008. "Statistical Discrimination with Peer Effects: Can Integration Eliminate Negative Stereotypes?" *Review of Economic Studies* 75:579–596.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5): 1220–1240.

Collins, Dave. 2018. "'Predictive Policing': Big-City Departments Face Lawsuits." *AP.* `https://apnews.com/b11e4bca11e548d3af7a63f24e348c6f`.

Dhami, Sanjit. 2016. *The Foundations of Behavioral Economic Analysis.* Oxford, UK: Oxford University Press.

Doleac, Jennifer L., and Luke C.D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." *The Economic Journal* 123:F469–F492.

Eckhouse, Laurel. 2019. "Everyday Risk: Disparate Exposure and Racial Inequality in Police Violence." Manuscript.

Epp, Charles R., Steven Maynard-Moody, and Donald P. Haider-Markel. 2014. *Pulled Over: How Police Stops Define Race and Citizenship.* Chicago, IL: University of Chicago Press.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. “Statistical Discrimination or Prejudice? A Large Sample Field Experiment.” *Review of Economics and Statistics* 96 (1): 119–134.

Eyster, Erik, and Matthew Rabin. 2005. “Cursed Equilibrium.” *Econometrica* 73 (5): 1623–1672.

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.” *Journal of the American Statistical Associationt* 102 (479): 813–823.

Glaser, Jack. 2006. “The efficacy and effect of racial profiling: A mathematical simulation approach.” *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 25 (2): 395–416.

———. 2015. *Suspect Race: Causes and Consequences of Racial Profiling.* New York: Oxford University Press.

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. “Precinct or Prejudice? Understanding Racial Disparities in New York City’s Stop-and-Frisk Policy.” *The Annals of Applied Statistics* 10 (1): 365–394.

Harcourt, Bernard E. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* Chicago, IL: University of Chicago Press.

Kahneman, Daniel, and Amos Tversky. 1973. “On the Psychology of Prediction.” *Psychological Review* 80 (4): 237–251.

Katz, Irwin. 1991. “Gordon Allport’s “The Nature of Prejudice”.” *Political Psychology* 12 (1): 125–157.

Knowles, John, Nicola Persico, and Petra Todd. 2001. “Racial Bias in Motor Vehicle Searches: Theory and Evidence.” *Journal of Political Economy* 109 (1): 203–229.

Knox, Dean, Will Lowe, and Jonathan Mummolo. 2019. "The Bias Is Built In: How Administrative Records Mask Racially Biased Policing." Manuscript. `https://ssrn.com/abstract=3336338`.

Knox, Dean, and Jonathan Mummolo. 2020. "Making Inferences about Racial Disparities in Police Violence." *PNAS* 117 (3): 1261–1262.

Komisarchik, Mayya, Maya Sen, and Yamil R. Velez. 2019. "The Political Consequences of Ethnically Targeted Incarceration: Evidence from Japanese-American Internment During WWII." Manuscript. `http://j.mp/2B0YNBG`.

Lerman, Amy E., and Vesla Weaver. 2014. "Staying out of Sight? Concentrated Policing and Local Political Action." *The Annals of the American Academy of Political and Social Science:* 202–219.

Levy, Gilat, and Ronny Razin. 2015. "Correlation Neglect, Voting Behavior, and Information Aggregation." *American Economic Review* 105 (4): 1634–1645.

Lum, Kristian, and William Isaac. 2016. "To Predict and Serve?" *Significance:* 15–19.

McCall, Andrew. 2019. "Resident Assistance, Police Chief Learning, and the Persistence of Aggressive Policing Tactics in Black Neighborhoods." *Journal of Politics* 81 (3): 1133–1142.

Minozzi, William. 2013. "Endogenous Beliefs in Models of Politics." *American Journal of Political Science* 57 (3): 566–581.

Nathan, Noah L., and Ariel White. Forthcoming. "Experiments on and with Street-Level Bureaucrats." In *Handbook of Advances in Experimental Political Science,* edited by James Druckman and Donald Green. Cambridge University Press.

O'Flaherty, Brendan, and Rajiv Sethi. 2019. *Shadows of Doubt: Stereotypes, Crime, and the Pursuit of Justice.* Cambridge, MA: Harvard University Press.

Ogden, Benjamin. 2019. “The Imperfect Beliefs Voting Model.” Manuscript. https://ssrn.com/abstract=2431447.

Patty, John W., and Roberto A. Weber. 2007. “Letting the Good Times Roll: A Theory of Voter inference and Experimental Evidence.” *Public Choice* 130 (3-4): 293–310.

Persico, Nicola. 2009. “Racial Profiling? Detecting Bias Using Statistical Evidence.” *Annual Review of Economics,* no. 1: 229–253.

Peyton, Kyle, Michael Sierra-Arévalo, and David G. Rand. 2019. “A Field Experiment on Community Policing and Police Legitimacy.” *PNAS* 116 (40): 19894–19898.

Phelps, Edmund S. 1972. “The Statistical Theory of Racism and Sexism.” *American Economic Review* 62 (4): 659–661.

Soss, Joe, and Vesla Weaver. 2017. “Police Are Our Government: Politics, Political Science, and the Policing of Race-Class Subjugated Communities.” *Annual Review of Political Science* 20:565–591.

The Sentencing Project. 2015. “Black Lives Matter: Eliminating Racial Inequity in the Criminal Justice System.” https://www.sentencingproject.org/wp-content/uploads/2015/11/Black-Lives-Matter.pdf.

Walker, Hannah L. 2020. “Targeted: The Mobilizing Effect of Perceptions of Unfair Policing Practices.” *The Journal of Politics* 82 (1).

Weaver, Vesla M., and Amy E. Lerman. 2010. “Political Consequences of the Carceral State.” *American Political Science Review* 104 (4): 817–833.