A Behavioral Theory of Discrimination in Policing*

Ryan Hübert† Andrew T. Little‡

April 2021

Abstract

A large economic literature studies whether racial disparities in policing are explained by animus or by beliefs about group crime rates. But what if these beliefs are incorrect? We analyze a model where officers form beliefs using crime statistics, but don't properly account for the fact that they will detect more crime in more heavily policed communities. This creates a feedback loop where officers over-police groups that they (incorrectly) believe exhibit high crime rates. This inferential mistake 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, misspecified models

*Authors are listed in alphabetical order.

†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, Black residents 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?[1] 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 makes less sense to think of the belief and preference channels as distinct and separable causes of discrimination. Instead, they may be fundamentally intertwined.

This is not just a hypothetical. Even highly-trained researchers make mistakes in interpreting crime data (Heckman and Durlauf 2020; Knox, Lowe, and Mummolo 2020; Knox and Mummolo 2020), and there is little reason to think that the relevant

[1] As discussed in more detail below, two recent papers consider this possibility and provide empirical tests, though not in the context of policing (Bohren et al. 2019; Bohren, Imas, and Rosenberg 2019).

decision-makers will generally do better (see, e.g., Glaser 2015, for an overview). In other words, there is good reason to think that the relevant decision-makers have a *misspecified* model of how their behavior and underlying crime rates map to crimes caught (Esponda and Pouzo 2016; Heidhues, Koszegi, and Strack 2018; Bohren 2016).

Having misspecified models is particularly important in this context, as 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 bad data or don't interpret it correctly, 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. 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. These officers' model of the world is misspsecified in the sense that they don't correctly account 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. 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. To

make the effect of incorrect belief formation clear, we set up the model such that when each officer has correct beliefs, 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 that officer's 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.[2] 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.

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

1 Explaining Policing Disparities and Incorrect Beliefs

Policing disparities have been well documented.[3] And, there is convincing evidence that statistical estimates of policing disparities may actually understate the extent of those disparities (Knox, Lowe, and Mummolo 2020; 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)[4] 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 related domains like labor markets 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; Harris, Ash, and Fagan 2020; Nathan and White, 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.[5] For policymakers and police departments seeking to reduce polic-

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

[4]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.

[5]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."

ing disparities, it matters why those disparities exist. 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 optimal to police those groups with different intensities. We capture this in our models by allowing for the possibility that true crime rates are different between groups.

A key question for explanations focused on statistical discrimination is whether decision-makers' beliefs about relative crime rates match reality. Substantial empirical evidence suggests that the relevant decision-makers may not have accurate beliefs about the key facts that drive policing allocations (see, for example, Glaser 2015). Cast in the language of a burgeoning literature, they may have "misspecified" models of the data generating process for crime (Bohren 2016; Esponda and Pouzo 2016; Heidhues, Koszegi, and Strack 2020; Levy, Razin, and Young 2020).[6] We build on a strand within this literature on how incorrect beliefs and behavior

[6]Of course, a core contention of much research on discrimination is that even "correct" statistical reasoning 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).

can interact (see Esponda and Pouzo 2016, for a general analysis of such games). For example, Heidhues, Koszegi, and Strack (2018) show that overconfidence may cause decision-makers to form incorrect beliefs about other aspects of the world (e.g., the ability of their subordinates), which in turn changes their behavior and future inferences, potentially leading to large distortions of both. In a similar vein, Levy, Razin, and Young (2020) study a model of political competition where some voters have a misspecified model of how policies map to outcomes, and learn from the equilibrium choices made by politicians.

Allowing for misspecified models has particularly important implications for the study of discrimination.[7] 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 may be based on biased or otherwise incorrect beliefs. This paper also highlights the challenge of distinguishing between the different causes of discrimination, particularly when incorrect beliefs are possible. Bohren, Imas, and Rosenberg (2019) derive an empirical test for this purpose using dynamic data, and find evidence that discrimination is driven by incorrect beliefs in the context of an online Q&A forum. We apply these ideas to the salient context of policing by micro-founding a particular kind of misspecified model that is prevalent in policing. We demonstrate how taste-based discrimination by police officers is endogenously exacerbated by incorrect beliefs, but our findings could also apply to discrimination in other domains.

More generally, our work also contributes to a broader literature in behavioral economics that focuses on the ways in which decision makers deviate from standard assumptions in formal analysis (for a comprehensive overview, see Dhami 2016). In particular, we add to a growing body of theoretical political economy models which explore the implications of non-standard belief formation (for example, Benabou and Tirole 2006; Levy and Razin 2015; Minozzi 2013; Acharya, Blackwell, and Sen 2018; Ogden 2019; Patty and Weber 2007). We also build on work which examines how decision-makers may overweight or underweight information from

[7]Heidhues, Koszegi, and Strack (2020) study how overconfidence can lead to inaccurate beliefs about groups which could lead to discriminatory behavior, but do not model the behavior itself (nor feedback loops between behavior and beliefs).

sources like their prior beliefs (Kahneman and Tversky 1973; Camerer 1995) or the decisions made by others (Eyster and Rabin 2005; Jehiel 2018) or by ones' past self (Benabou and Tirole 2006). 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*.

In terms of our substantive application, 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 in the domain we study, 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.[8] 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.

2 Model of a Single 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

[8]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.

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.[9]

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 can't choose to allocate *all* of his time to one group or the other. Formally, 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 policing only one group is infeasible—reduce the number of cases to consider for some of our results.[10]

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 decides how to allocate time between neighborhoods, and not between social groups.

[9] 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 for this would add complexity to the model without obviously changing our results.

[10] 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.

We assume that the allocation of policing effort, as reflected by w_A, affects the detection of crime. As a result, our model more directly captures "proactive policing," rather than "reactive policing" where officers respond to reports of crimes in progress or which have already occurred (e.g., via 911 calls). The model is also 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 the 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 addition to these key assumptions, we place several relatively mild technical assumptions on the officer utility:

Assumption 1. *The officer utility is:*

$$u(t_A c_A, t_B c_B) \tag{1}$$

where $t_J > 0$, and the utility function $u(x_1, x_2)$ is (i) symmetric in the two arguments ($u(x_1, x_2) = u(x_2, x_1)$), (ii) continuously differentiable, (iii) strictly increas-

ing and concave in both arguments ($u_1 > 0$, $u_{11} < 0$, $u_2 > 0$, $u_{22} < 0$), additively separable ($u_{12} = 0$), and (iv) homogeneous with positive degree.

The t_A and t_B terms represent the officer's "taste" for catching crimes among group A and B, respectively. Given part (iii) of the assumption, higher values of t_J will make the officer value catching crimes among group J more.[11]

The remaining assumptions are for technical convenience. The symmetry assumption implies that the labeling of the groups is arbitrary, in the sense that the only differences between them are captured by the t_J and p_J parameters. Additive separability allows us to focus on the direct effect where catching more crime among group J decreases the marginal return to catching more crime among this group. We thus set aside potential indirect effects where this also means allocating less time to $-J$, which can affect the return to policing group J through the cross-partial derivative. As the proof of Lemma 1 shows, it is sufficient that this cross-partial is not too positive or too negative (relative to the concavity in each argument). The homogeneity assumption is primarily to provide a convenient characterization of the optimal policing allocation.

If the officer has correct beliefs about the p_J parameters, then the optimal allocation of time is a straightforward maximization of (1). When the officer has correct beliefs, we say he has *full information*.

Lemma 1. *Let $r_t = t_A/t_B$ and $r_p = p_A/p_B$. Given Assumption 1, if the officer knows r_p then there is a unique w_A which optimizes (1), which we write as $w_A^{br}(r_t, r_p)$.*

(i) w_A^{br} is increasing in both arguments, and where w_A^{br} is interior, strictly increasing in both arguments.

(ii) $w_A^{br}(1, 1) = 1/2$.

Proof See the appendix.

We will also use $w_A^{\dagger} = w_A^{\text{br}}(r_t, r_p)$ to refer to the officer allocation with full information.

[11] 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*.

Moving forward, we will describe how changes in these ratios affect behavior rather than the primitive p_J and t_J terms. This is particularly convenient as these two parameters correspond exactly to the standard explanations for discrimination.

The r_t parameter reflects the preference for catching crimes among group A, relative to group B, which captures the possibility for taste-based discrimination. We say that if $r_t > 1$ the officer has *animus* towards group A, and $r_t < 1$ indicates animus towards group B. The r_p parameter reflects the *true* ratio of the two groups' crime rates, capturing the possibility for 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.

Part (ii) of the lemma states that in the absence of either of these forces—i.e., $r_t = 1$ means the officer has no animus towards either group and $r_p = 1$ means the crime rates are equal—then the officer splits his time equally between policing the two groups.

Importantly, all of our following results will hold for any utility function with the properties of Lemma 1. So, for example, the officer does not necessarily need to be motivated by maximizing the amount of crime caught (Stashko 2020); what really matters is that they want to allocate more time policing groups with higher crime rates, as well as groups against which they have animus.

To reduce the cases to consider, we place one more assumption on the utility function which ensures an interior solution with full information:

Assumption 2.

$$\frac{u_2(\underline{w}, 1-\underline{w})}{u_1(1-\underline{w}, \underline{w})} < r_t r_p < \frac{u_2(\overline{w}, 1-\overline{w})}{u_1(1-\overline{w}, \overline{w})}$$

In words, this states that at the lower bound $\underline{w}$ the marginal return to policing group A is higher than the marginal return to group B, and at the upper bound the reverse is true.

Lemma 2. *Given Assumptions 1 and 2, the full information allocation is interior:* $w_A^\dagger \in (\underline{w}, \overline{w})$.

As we will show, this does not preclude a corner solution when we allow the officer to have incorrect beliefs. That is, to highlight the effect of incorrect belief formation, we want to start at a benchmark where the officer chooses to spend substantial time policing both groups.

Main Example For illustrations, we will use the following utility function which meets assumption 1 and leads to tidy closed form solutions:

$$u(c_A, c_B) = \sqrt{t_A c_A} + \sqrt{t_B c_B} = \sqrt{t_A p_A w_A} + \sqrt{t_B p_B (1 - w_A)} \tag{2}$$

With this utility, the officer optimal allocation as a function of the ratios is:

$$w_A^\dagger = \frac{r_t r_p}{1 + r_t r_p} \tag{3}$$

Note that since $r_t > 0$ and $r_p > 0$, this is always strictly between 0 and 1, and hence as long as $\underline{w}$ and $\overline{w}$ are sufficiently close to 0 and 1, then $w_A^\dagger$ is interior.[12]

Disparities 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 $w_A^{\text{stat}} = w_A^{\text{br}}(1, r_p)$ to be the "statistical policing" allocation, which reflects what

[12]This follows from the fact that with the square root function the marginal return to policing group J approaches infinity as $c_J \to 0$.

an officer does if he has no animus toward either group but statistically discriminates based on differences in the (true) crime rates. Following Bohren et al. (2019), we refer to $w_A^{\text{stat}} - 1/2$ as the "traditional statistical discrimination," which will contrast with the "inaccurate statistical discrimination" which arises when the officer does not know r_p. The difference between what the officer chooses and this statistical benchmark, $w_A^\dagger - w_A^{\text{stat}}$, represents taste-based discrimination. Taken together, the policing disparity when the officer has full information can be decomposed as follows:[13]

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

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 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, heavily policing certain social groups can have other spillovers, such as reducing political participation (Weaver and Lerman 2010) and causing some communities' interactions with the state to overwhelmingly consist of negative interactions with the police (Soss and Weaver 2017).

[13]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.

2.1 Policing with a Misspecified Model

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.

In general, we assume that in the belief formation stage, the officer may have a misspecified model, in the sense that he misunderstands how policing allocations affect the crime detected. In the extreme, the officer might infer that the relative number of crimes caught among the groups is the same as the relative crime rates. As this involves the officer forming a posterior belief without properly conditioning on all relevant information, it is *non-conditioning bias*.

Formally, suppose at the stage where the officer is forming beliefs about the relative crime rates, he does so as if he thinks the crime detection function is $\tilde{c}(w_J, p_J)$, which may not match the real function $c(w_J, p_J)$.[14] In our main results below where the officer has a misspecified model, we will maintain the assumption from above that crimes are detected according to the function $c(w_J, p_J) = w_J p_J$. However, before getting to this, it is instructive to consider how this kind of misspecified model affects belief formation for a general $c(w_J, p_J)$ with the weaker assumption that it is continuously differentiable, weakly increasing in w_J, and strictly increasing in

[14]There is a tension in literally interpreting this source of misspecification since the officer behaves as if he does know the correct crime function when solving the optimization problem for his allocation, but not when forming beliefs. The simplest resolution, which will be more natural in the model with multiple officers, is that the officer thinks about crime detection differently when choosing his own allocation versus when he forms beliefs based on crime data, which in reality also depends on the choices made by others. Alternatively, what matters for the allocation stage is not the specific functional form but the fact that the optimal allocation is increasing in r_t and r_p. Similarly, what matters in the belief formation stage is not that incorrect beliefs are driven by a misspecified c function, but that beliefs become a function of the allocation.

p_J. Further, assume that the officer's belief $\tilde{c}$, while potentially not equal to c, still shares these properties.

This assumption implies that, upon observing a crime level c_J, the officer will infer that the crime rate of this group is the value of p_J that solves $c_J = \tilde{c}(w_J, p_J)$. An important implicit assumption here is that the officer observes "enough data" that the crime detected is exactly $c(w_J, p_J)$. That is, we do not explicitly model the randomness inherent to the process. We do so to keep the model simple and the focus on the application; as elaborated when we introduce our full solution concept below, several theoretical papers study the convergence of beliefs and actions when explicitly modeling such randomness.

If $\hat{p}(c_J, w_J)$ is the value of p_J that solves $c_J = \tilde{c}(w_J, p_J)$ and $c_J = c(w_J, p_J)$, then the officer's inference about the crime rate of group J, which we denote $\tilde{p}_J$, is $\tilde{p}_J = \hat{p}(c(w_J, p_J), w_J)$. We are primarily interested in when there is an interaction between the officer choice w_J and this resulting belief. Fortunately, there is a clean characterization of when such interactions occur.

Proposition 3. *The officer's belief about the crime rate of group J is strictly increasing in w_J if he strictly underestimates the impact of w_J on c_J, and is strictly decreasing in w_J if he strictly overestimates this quantity:*

$$\textit{sign}\left(\frac{\partial \tilde{p}_J}{\partial w_J}\right) = \textit{sign}\left(\frac{\partial c}{\partial w_J} - \frac{\partial \tilde{c}}{\partial w_J}\right)$$

We now return to our original assumption that $c(w_J, p_J) = w_J p_J$. To simplify moving forward, we use a functional form for $\tilde{c}$ that clarifies exactly how the officer in our model underestimates the effect of his policing allocation:

$$\tilde{c}(w_J, p_J) = (1 - \nu) p_J w_J + \nu p_J$$

That is, his (potentially misspecified) model of crime detection is a weighted average of the *true* amount of crime detected and the crime rate ignoring his own policing efforts. The weight $\nu \in [0, 1]$ specifies the severity of his non-conditioning bias. So, as $\nu \to 0$ the officer has a correct understanding of how crime works. If

$\nu > 0$ he underestimates the impact of his action on the amount of crime caught.

If the officer has this model of crime detection in his head, then after observing a crime rate c_J and his own policing intensity w_J, his (possibly distorted) belief $\tilde{p}_J$ solves:

$$c_J = (1-\nu)\tilde{p}_J w_J + \nu\tilde{p}_J = \tilde{c}(w_J, p_J)$$

Since $c_J = c(w_J, p_J) = w_J p_J$, we can substitute and rearrange:

$$\tilde{p}_J = \frac{w_J p_J}{(1-\nu)w_J + \nu}$$

Again note that if $\nu = 0$, this simplifies to p_J, and is unaffected by w_J. However, for any $\nu > 0$, the belief will increase in w_J.

Combining the group ratios the belief about the ratio is:[15]

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

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.

Regardless of the specific mechanism that generates this belief, our assumption that police officers exhibit non-conditioning bias is not outlandish. Recent studies provide causal evidence that experimental subjects neglect selection effects and thus make faulty inferences about a state of the world (e.g., Barron, Huck, and Jehiel 2019; Enke 2020). Several examples suggest the phenomenon extends to the real-world context of policing. 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

[15] Algebraically, our bias ends up resembling a technology used in Benabou and Tirole (2006), who use it to model how individuals bias their future beliefs by limiting recall of particular kinds of information and not fully adjusting for this limited recall.

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.[16]

Equilibrium We have characterized how the officer's 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 1. *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) $\tilde{r}_p^* = \tilde{r}_p(w_A^*)$.

If $\left.\frac{\partial w_A^{br}}{\partial w_A}\right|_{w_A = w_A^*} < 1$*, we say the equilibrium is **stable**.*

The two equilibrium conditions state that the officer choice is optimal given his beliefs, and that his beliefs are given by equation (4). The stability condition ensures that a perturbation to the officer choice would lead back to the equilibrium, as illustrated below.

[16] 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 $\nu \in [0, 1]$.

This is similar to what Esponda and Pouzo (2016) call a "Berk Nash Equilibrium." Much of the theoretical literature on misspecified models provides general conditions under which beliefs and behavior do in fact converge to such a stable point (Esponda and Pouzo 2016; Bohren 2016; Levy, Razin, and Young 2020). To keep the focus on our application, we will analyze behavior at a stable point.

Our main technical result is that an equilibrium always exists, and with a stronger assumption, is unique.

Proposition 4. *A stable equilibrium exists in the single officer model. If ν is sufficiently small, the equilibrium is unique.*

The condition for uniqueness is that the officer's bias is not too large, as this limits the complementarity between his action and belief.

Main Example (continued) This condition on ν is not always necessary; in fact, for our main example with a utility function given by (2), there is a unique equilibrium in which the officer chooses a policing allocation

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

where

$$\widehat{w}_A = w_A^\dagger + \frac{\nu(r_t r_p - 1)}{(1-\nu)(1 + r_t r_p)} \tag{5}$$

This is because there is a unique solution to $w_A = \frac{r_t \tilde{r}_p(w_A)}{1 + r_t \tilde{r}_p(w_A)}$ given by $\widehat{w}_A$. If $\widehat{w}_A$ lies in $[\underline{w}, \overline{w}]$, then it corresponds to an equilibrium allocation. Whenever $\widehat{w}_A$ does not lie in $[\underline{w}, \overline{w}]$, then there is an equilibrium at a corner solution.[17]

To allow for clean statements about how inaccurate beliefs affect equilibrium behavior, our remaining technical results (and illustrations) focus on the case where there is a unique equilibrium.

[17] We formally state this equilibrium in Proposition 8 in the appendix.

Illustrations 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 (4) 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.

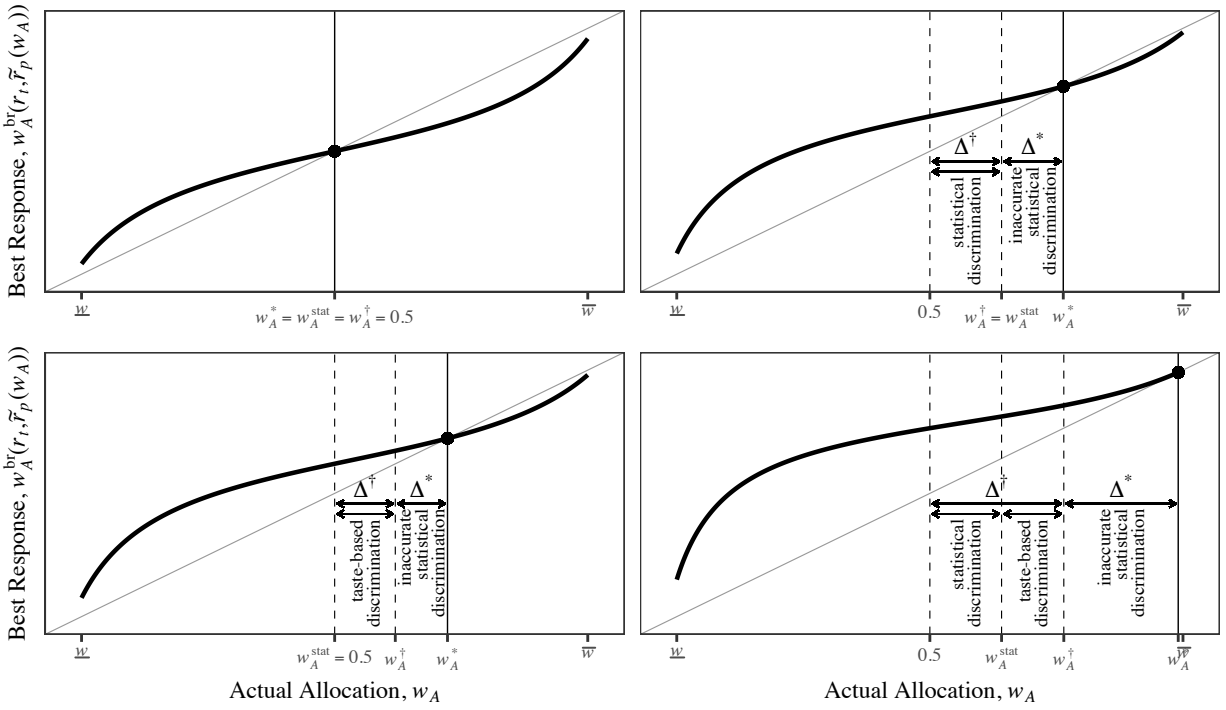
Figure 1 illustrates this dynamic process for our main example. In each panel (which vary in their values of r_t and r_p), the black curves trace out $w_A^{\text{br}}(r_t, \tilde{r}_p(w_A))$ as a function of w_A. The grey 45-degree line represents points where the best response allocation equals the 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 necessary condition for an equilibrium is to lie 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. Further, these diagrams also illustrate the stability condition.[18] 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

[18] 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 inaccurate statistical 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 ($r_t > 1$).



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 $w_A^\dagger > 1/2$ (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 the full information benchmark 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. Following (Bohren et al. 2019), we refer to it as *inaccurate statistical discrimination*. We will show below that inaccurate statistical 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.

Inaccurate statistical 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 roughly doubles the policing disparity relative to the full information benchmark. In a sense, this is all statistical discrimination, but roughly half of it is driven by false beliefs.

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 a corner solution even though his policing allocation would be interior

if he had full information. As demonstrated below, such corner solutions do 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. When conceptualized in this way, our model shows that taste-based discrimination can *cause* (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.

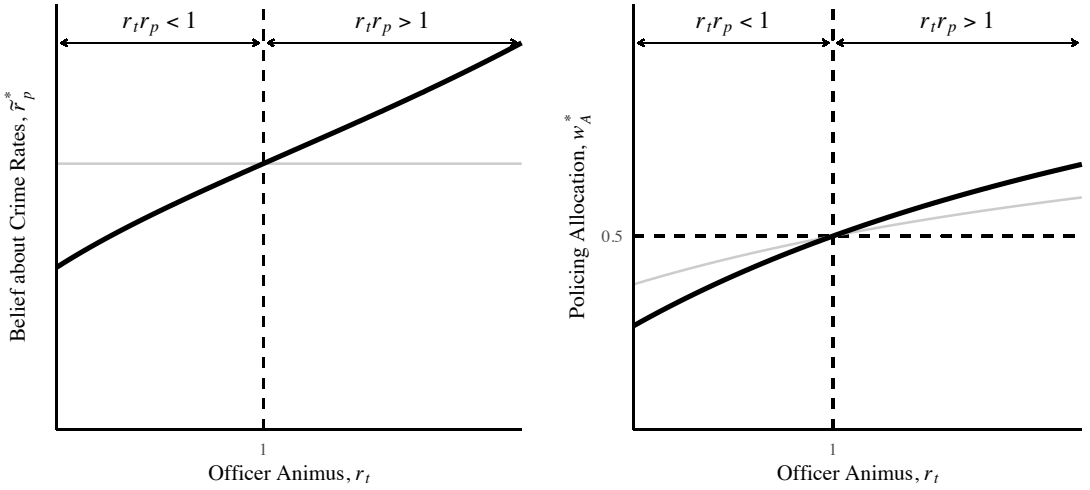
To be more concrete about how this works in our model, consider the following. First, the officer's animus causes him to allocate more policing effort toward one group. Then, since he spends more time policing that group, he sees more crimes among members of that group. Finally, his non-conditioning 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 inaccurate statistical 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 the full information benchmark (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 an-

imus 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 in the full information benchmark. 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 with the full information benchmark, as the right panel shows.

Figure 2: The solid grey lines depicts the officer's policing allocation and belief about relative crime rates with the full information benchmark, and the solid black lines depicts his policing allocation and belief about relative crime rates when he polices with a misspecified model. 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.



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.

The previous analysis suggests that inaccurate statistical 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 having a misspecified model of crime does not amplify policing disparities.[19] If there would be no policing disparity if the officer had full information, then non-conditioning bias does not amplify policing disparities. Formally, this occurs if $r_t r_p = 1$ so that $w_A^* = w_A^\dagger = 1/2$. In this case, the officer's misspecified model 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 equal across groups ($r_p = 1$). 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 $r_p = 1/r_t$.

However, this is a very specific situation. If $r_t r_p < 1$, 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 $r_t r_p > 1$, 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 when the officer polices with a misspecified model, it generically amplifies whatever disparities would exist if officer formed accurate beliefs about crime rates.

Proposition 5. *For any $\nu \in (0, 1)$:*

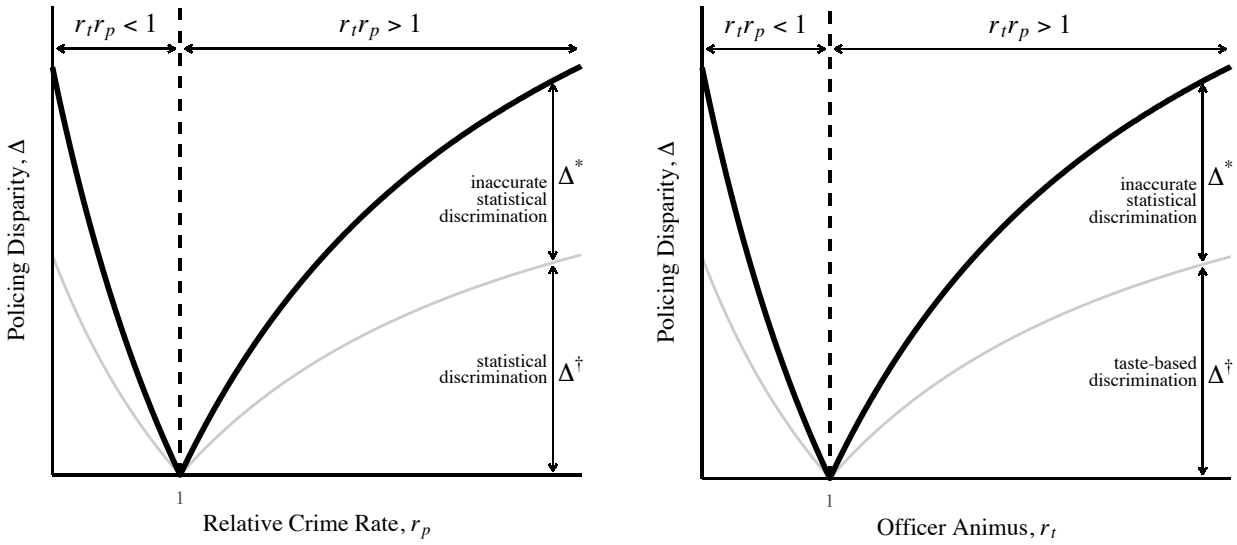
(i) If $r_t r_p = 1$, then an there is an equilibrium with no policing disparity (since $w_A^* = w_A^\dagger = 1/2$), and the officer has correct beliefs about crime, $\tilde{r}_p^* = r_p$.

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

If the officer's policing allocation is not at a corner ($\underline{w}$ or $\overline{w}$), then the disparity caused by inaccurate statistical discrimination is strictly positive as as $r_t r_p$ 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

[19] If we relax Assumption 2, then there is a second situation in which there is no inaccurate statistical discrimination. If the officer would engage in extreme policing regardless whether he has a non-conditioning bias, then trivially, incorrect beliefs do not increase the policing disparity.

Figure 3: In each panel, we plot the policing disparity that emerges in an interior 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 $r_t r_p \neq 1$, the officer always engages in either statistical or taste-based discrimination, *as well as* inaccurate statistical discrimination.



and the black line depicts the entire policing disparity. Note that in either panel, as long as $r_t r_p \neq 1$, then inaccurate statistical discrimination causes the policing disparity to be higher than it otherwise would have been with only taste-based and statistical discrimination.

For Proposition 5 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 whenever there is an interior allocation. In some situations, such as for our main example given by the utility function (2), the policing disparity increases until the officer reaches a corner solution.

Proposition 6. *Assume $r_t r_p \neq 1$. Then if there is a unique equilibrium with an interior solution, the policing disparity caused by inaccurate statistical discrimination (Δ^*) is strictly increasing in ν.*

Figure 4: As the severity of the officer's inferential mistake (i.e., ν) increases, so does the policing disparity caused by inaccurate statistical discrimination.

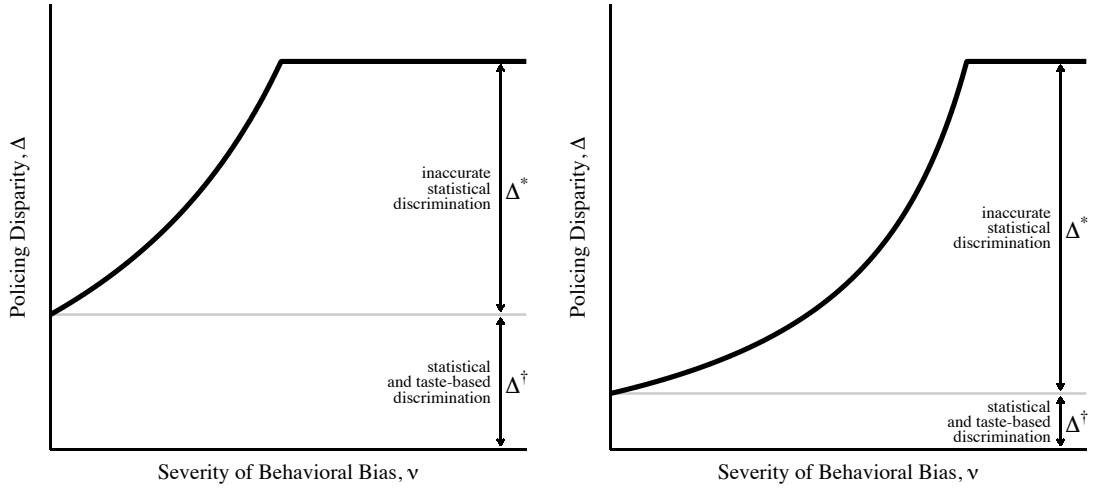


Figure 4 depicts the result for two situations. In each panel, the x-axis plots ν and the y-axis plots the policing disparity. The left panel depicts a situation where statistical and taste-based discrimination cause a large disparity, and the right panel depicts a situation where statistical and taste-based discrimination cause a small disparity. In either case, at any interior allocation inaccurate statistical discrimination strictly increases as ν increases. Further, for these parameters, when ν gets sufficiently high the officer eventually starts policing one group as much as possible, i.e., the equilibrium is at a corner solution.

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. In other words, policing with a misspecified model about crime does not preclude the possibility of a stable equilibrium of our model.

Second, policing with a misspecified model of the type we study generically causes the officer to engage in inaccurate statistical discrimination in which he over-polices 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. The term "prejudice" in prior research and in 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 inaccurate statistical discrimination amplifies the policing disparity caused by taste-based and/or statistical discrimination. In particular, if $\nu > 0$ and $r_t r_p \neq 1$, then he will always overpolice one group more than he otherwise would with full information. Moreover, the extent to which inaccurate statistical discrimination amplifies existing a policing disparity increases with the severity of his inferential mistake. As a result, policing with a misspecified model can sometimes dramatically increase policing disparities.

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.

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.

To simplify, we assume each officer cares only about the number of crimes that he catches, and use the utility function of our main example in the previous section:

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

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 an interior solution, the best response for each officer i depends on his animus ($r_{t,i}$) and the true ratio of crime rates (r_p):

$$w^{\dagger}_{A,i} = \frac{r_{t,i} r_p}{1 + r_{t,i} r_p}.$$

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:

$$\tilde{r}_{p,i}(w_A) = \frac{\frac{c_A}{\nu_i + (1-\nu_i) w_{A,1} + \nu_i + (1-\nu_i) w_{A,2}}}{\frac{c_B}{\nu_i + (1-\nu_i) w_{B,1} + \nu_i + (1-\nu_i) w_{B,2}}} = \frac{\frac{c_A}{2\nu_i + (1-\nu_i) w_A}}{\frac{c_B}{2\nu_i + (1-\nu_i)(2 - w_A)}} \tag{6}$$

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, 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:

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

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 2. *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 (6) evaluated at* w^*_A.

If a condition analogous to the single officer model is met (see Appendix B), we say the equilibrium is stable.

(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), 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 7. *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.

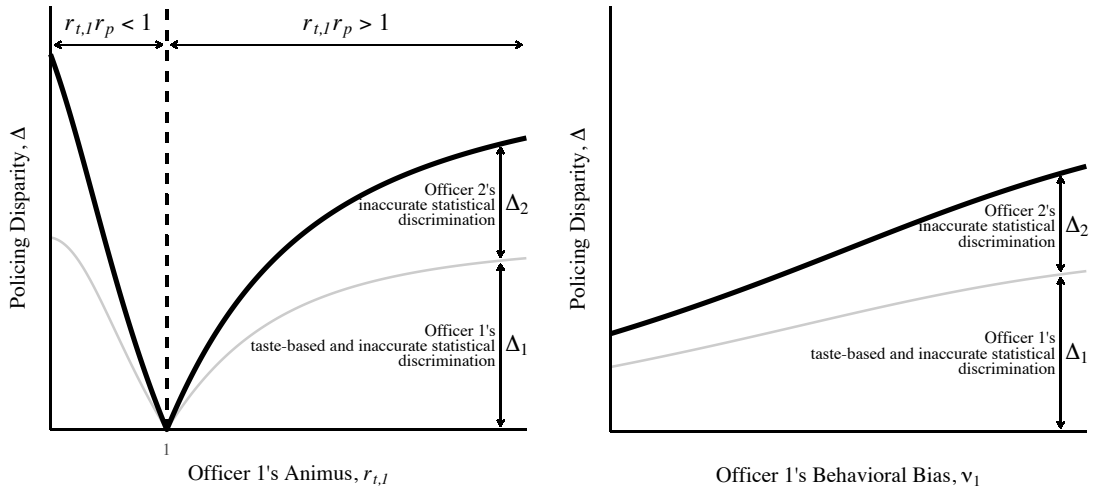
Finally, we illustrate how this affects individual and aggregate discrimination. To consider each officer's discrimination separately, define the policing disparities generated by each officer as follows:

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

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 animus (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 inaccurate statistical 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 inaccurate statistical 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 these quantities as a function of $r_{t,1}$.

Since officer 2 has no animus and crime rates are equal, he would not discriminate if he had full information, even if officer 1 did. 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*. As a result, the officers in our model police with a misspecified model.

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. This kind of discrimination amplifies 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. Our model thus shows how racial animus and discrimination based on incorrect beliefs are intertwined.

We also extend the model to examine how policing with a misspecified model can generate discrimination when there are multiple officers. 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.

References

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

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.

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/`.

Barron, Kai, Steffen Huck, and Philippe Jehiel. 2019. "Everyday Econometricians: Selection Neglect and Overoptimism When Learning from Others." Manuscript.

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.

Benabou, Roland, and Jean Tirole. 2006. "Belief in a just world and redistributive politics." *The Quarterly journal of economics* 121 (2): 699–746.

Bohren, J. Aislinn. 2016. "Informational herding with model misspecification." *Journal of Economic Theory* 163:222–247.

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.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The dynamics of discrimination: Theory and evidence." *American economic review* 109 (10): 3395–3436.

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.

Enke, Benjamin. 2020. "What You See Is All There Is." *Quarterly Journal of Economics* 135 (3): 1363–1398.

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.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica* 84 (3): 1093–1130.

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.

Harris, Allison P., Elliott Ash, and Jeffrey Fagan. 2020. "Fiscal Pressures and Discriminatory Policing: Evidence from Traffic Stops in Missouri." *Journal of Race, Ethnicity, and Politics.* https://doi.org/10.1017/rep.2020.10.

Heckman, James J., and Steven N. Durlauf. 2020. "Comment on "An Empirical Analysis of Racial Differences in Police Use of Force" by Roland G. Fryer Jr." *Journal of Political Economy* 0 (ja): null. doi:10.1086/710976. eprint: https://doi.org/10.1086/710976. https://doi.org/10.1086/710976.

Heidhues, Paul, Botond Koszegi, and Philipp Strack. 2018. "Unrealistic expectations and misguided learning." *Econometrica* 86 (4): 1159–1214.

———. 2020. "Overconfidence and Discriminiation." Manuscript.

Jehiel, Philippe. 2018. "Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism." *American Economic Review* 108 (6): 1582–1597.

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

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. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* 114 (3): 619–637.

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.

Levy, Gilat, Ronny Razin, and Alwyn Young. 2020. "Misspecifed Politics and the Recurrence of Populism." Manuscript.

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.

Stashko, Allison. 2020. "Do Police Maximize Arrests or Minimize Crime? Evidence from Racial Profiling in U.S. Cities." Manuscript.

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.

Online Appendix

"A Behavioral Theory of Discrimination in Policing"

Ryan Hübert
UC Davis

Andrew T. Little
UC Berkeley

April 2021

FOR ONLINE PUBLICATION

A Racial Profiling and the Geography of Policing

The analysis in the main text assumes that police officers decide to allocate their time between policing two groups of people. In the United States, the prevailing law is unclear about whether such group-based profiling is permissible (for an extended discussion, see Knowles, Persico, and Todd 2001). However, under the U.S. Constitution, policies that explicitly treat members of protected categories differently are subject to strict scrutiny (see *Brown v. Board of Education of Topeka*, 1954). A policy of explicitly using group membership to allocate policing resources is not likely to survive a strict scrutiny legal analysis.

We focus on this simple, but potentially illegal, decision-making process in text because it allows us to more clearly focus on our core arguments. However, it can be microfounded with a more complex model where a police chief decides how many policing resources to devote to two neighborhoods: 1 and 2. Formally, assume he devotes n_1 of his time to policing neighborhood 1 and $n_2 = 1 - n_1$ of his time to policing neighborhood 2. Also assume that each neighborhood is comprised of members of the two groups, A and B. Within a neighborhood i, we assume that police interact with a member of group A with probability α_i and a member of group B with probability $1 - \alpha_i$. If police encounters with residents are random and iid, then one way to interpret α_i is that it represents the proportion of neighborhood i

that is comprised of members of group A. However, our flexible specification allows for the possibility that police come into contact with members of one group at a rate disproportionate to that group's share of the local population. (Although note that if α_i does not reflect the demographic makeup of neighborhood i, then we simply reintroduce concerns about racial profiling that motivate this microfoundation, just at a different point in the analysis.)

Conditional on a choice about how intensely to police each neighborhood, the share of group A individuals the police encounters is $\eta_A = n_1\alpha_1 + (1-n_1)\alpha_2 = \alpha_2 + (\alpha_1 - \alpha_2)n_1$ and the share of group B individuals the police encounters is $\eta_B = n_1(1-\alpha_1) + (1-n_1)(1-\alpha_2) = 1 - \eta_A$. Recall from the main text that w_A is defined as the share of time that the police officer devotes to policing group A, and $w_B = 1 - w_A$ is the corresponding share of time that the police officer devotes to policing group B. Then, η_A is equivalent to w_A and η_B is equivalent to w_B, and n_1 is a perfect proxy for w_A. More specifically, if police come into contact with group A more than group B in neighborhood 1 (alt. neighborhood 2), $\alpha_1 > \alpha_2$ (alt. $\alpha_1 < \alpha_2$), then increasing n_1 (alt. n_2) linearly increases w_A. Notice that in the extreme cases where $n_1 = 0$ and $n_1 = 1$, then $\eta_A = \alpha_2$ and $\eta_A = \alpha_1$, respectively. Then, α_1 and α_2 correspond the maximum and minimum possible allocations: $\underline{w} = \min\{\alpha_1, \alpha_2\}$ and $\overline{w} = \max\{\alpha_1, \alpha_2\}$.

In a model where police choose n_1 (and not w_A), the analysis in the main text is identical after substituting $\eta_A = \alpha_2 + (\alpha_1 - \alpha_2)n_1$ for w_A.

B Stability in the Multiple Officer Model

The first two equilibrium conditions for the two officer model can be combined as:

$$F_1(w_{A,1}, w_{A,2}) \equiv w_A^{\text{br}}(r_{t,1}, \tilde{r}_{p,1}(w_{A,1}, w_{A,2}, \nu_1)) - w_{A,1} = 0$$
$$F_2(w_{A,1}, w_{A,2}) \equiv w_A^{\text{br}}(r_{t,2}, \tilde{r}_{p,2}(w_{A,1}, w_{A,2}, \nu_2)) - w_{A,2} = 0$$

Cclose to an equilibrium, we want that for any "small" perturbation to both players' strategies, if the officers iteratively choose best responses given their new beliefs,

then the joint allocation would move back to the equilibrium. By standard results in the study of dynamic systems (e.g., Theorem 11.4 in Gintis 2009), this can be expressed by conditions on the matrix of the partial derivatives of the F_i functions:

Definition 3. *Let*

$$D(w_{A,1}, w_{A,2}) = \begin{bmatrix} \frac{\partial F_1}{\partial w_{A,1}} & \frac{\partial F_1}{\partial w_{A,2}} \\ \frac{\partial F_2}{\partial w_{A,1}} & \frac{\partial F_2}{\partial w_{A,2}} \end{bmatrix}.$$

an equilibrium in the two-officer model is is stable if:

(i) $tr(D(w^*_{A,1}, w^*_{A,2})) < 0$*, and*

(ii) $det(D(w^*_{A,1}, w^*_{A,2})) > 0$*.*

The first condition simplifies to

$$\left.\frac{\partial F_1}{\partial w_{A,1}}\right|_{w_A = w^*_A} + \left.\frac{\partial F_2}{\partial w_{A,1}}\right|_{w_A = w^*_A} < 0$$

Note that if both derivatives are negative (as required in the single officer model), this is always true.

The second condition becomes:

$$\left[\frac{\partial F_1}{\partial w_{A,1}}\frac{\partial F_2}{\partial w_{A,2}} - \frac{\partial F_1}{\partial w_{A,2}}\frac{\partial F_2}{\partial w_{A,1}}\right]_{w_A = w^*_A} > 0$$

To provide a more easily interpretable version of these conditions, define:

$$Y_i = \left.\frac{\partial w_A^{\text{br}}}{\partial \tilde{r}_{p,i}}\right|_{\tilde{r}_{p,i} = \tilde{r}_{p,i}(w^*_{A,1}, w^*_{A,2})}$$

$$Z_i = \left.\frac{\partial \tilde{r}_{p,i}(w_{A,1}, w_{A,2})}{\partial w_{A,1}}\right|_{w_A = w^*_A} = \left.\frac{\partial \tilde{r}_{p,i}(w_{A,1}, w_{A,2})}{\partial w_{A,2}}\right|_{w_A = w^*_A}.$$

Then:

$$\left.\frac{\partial F_i}{\partial w_{A,i}}\right|_{w_A = w^*_A} = (Y_i Z_i - 1) \qquad \left.\frac{\partial F_i}{\partial w_{A,-i}}\right|_{w_A = w^*_A} = Y_i Z_i$$

Plugging these into first stability condition gives:

$$(Y_1Z_1 - 1) + (Y_2Z_2 - 1) < 0 \iff Y_1Z_1 + Y_2Z_2 < 2 \tag{7}$$

and the second:

$$\begin{aligned}(Y_1Z_1 - 1)(Y_2Z_2 - 1) - (Y_1Z_1)(Y_2Z_2) > 0 \\ \iff Y_1Z_1 + Y_2Z_2 < 1\end{aligned}$$

which is stronger than condition (7) and hence the binding constraint.

An intuition for this condition is that due to the complementarities between action and belief, the deviations that are most apt not to return to an equilibrium are those where both officers increase or both officers decrease their allocations. And $Y_1Z_1 + Y_2Z_2$ is the marginal change in the best response as *both* officers increase their allocation to group A. So, this condition states that if both officers were to allocate slightly more time to group A or both allocated slightly less, their best responses would move back toward the equilibrium allocation.

C Proofs

Proof of Lemma 1 Since u is homogeneous with positive degree, for any α there exists a $k > 0$ such that:

$$u(\alpha t_A p_A w_A, \alpha t_B p_B (1 - w_A)) = \alpha^k u(t_A p_A w_A, t_B p_B (1 - w_A)) \tag{8}$$

Let $\alpha = 1/(t_B p_B)$, and note that w_A maximizes u if and only if it maximizes $(t_B p_B)^{-k} u$. Plugging this into equation (8) gives:

$$(t_B p_B)^{-k} u(t_A p_A w_A, t_B p_B (1 - w_A)) = u(r_t r_p w_A, 1 - w_A)$$

So, any interior w_A^{br} is characterized by the first order condition:

$$\frac{\partial u}{\partial w_A} = r_t r_p u_1(r_t r_p w_A, 1 - w_A) - u_2(r_t r_p w_A, 1 - w_A) = 0$$

The second derivative is

$$\begin{aligned}\frac{\partial^2 u}{\partial w_A^2} =& r_t r_p (r_t r_p u_{11}(r_t r_p w_A, 1-w_A) - u_{12}(r_t r_p w_A, 1-w_A)) \\ & - r_t r_p u_{12}(r_t r_p w_A, 1-w_A) + u_{22}(r_t r_p w_A, 1-w_A) < 0\end{aligned}$$

the u_{11} and u_{22} terms are strictly negative and the u_{12} are equal to zero by 1. If we loosen the assumption on u_{12}, the inequality holds as long as the cross-partial derivative is not too negative (relative to the u_{11} and u_{22} terms). The inequality holds a Since the objective function is globally strictly concave in w_A, and since it is continuous on a compact set, it must have a unique maximizer.

Since u is homogeneous degree k, u_1 is homogeneous degree $k-1$, so we can rewrite the first term of the FOC to give:

$$G(r_t, r_p, w_A) = (r_t r_p)^k u_1(w_A, (r_t r_p)^{-1}(1-w_A)) - u_2(r_t r_p w_A, 1-w_A) = 0 \quad (9)$$

Where w_A^{br} is interior, the change with respect to r_t is given by implicitly differentiating G

$$\frac{\partial w_A^{\text{br}}}{\partial r_t} = -\frac{\frac{\partial G}{\partial r_t}}{\frac{\partial G}{\partial w_A}}$$

The denominator is negative at any maximizer, and the numerator is:

$$\begin{aligned}\frac{\partial G}{\partial r_t} =& k r_p^k r_t^{k-1} u_1(w_A, (r_t r_p)^{-1}(1-w_A)) \\ & - (r_t r_p)^k (u_{12}(w_A, (r_t r_p)^{-1}(1-w_A)) r_t^{-2} - r_p w_A u_{12}(r_t r_p w_A, 1-w_A)\end{aligned}$$

The first term is strictly positive, and the second two drop out since $u_{12} = 0$. As long as u_{12} is not too positive, then $\frac{\partial w_A^{\text{br}}}{\partial r_t} > 0$. So, at any interior solution, the optimal allocation is strictly increasing in r_t, and since the FOC is strictly increasing in r_t the optimizer is weakly increasing in r_t even when there is a corner solution.

As r_t and r_p enter into the utility symmetrically, the proof for $\frac{\partial w_A^{\text{br}}}{\partial r_p} > 0$ follows an identical logic.

For part iii, $u(x,y) = u(y,x)$ implies $u_1(x,y) = u_2(y,x)$. The FOC when

$r_t r_p = 1$ is

$$u_1(w_A, 1 - w_A) = u_2(w_A, 1 - w_A)$$

which is clearly met at $w_A = 1/2$.

Proof of Lemma 2 The proof of Lemma 1 shows that the first derivative of the objective function is continuous and strictly decreasing in w_A. So, there will be an interior solution if and only if it is strictly positive at $w_A = \underline{w}$ and strictly negative at $w_A = \overline{w}$. The first condition requires:

$$\begin{aligned} r_t r_p u_1(r_t r_p \underline{w}, 1 - \underline{w}) &> u_2(r_t r_p \underline{w}, 1 - \underline{w}) \\ r_t r_p &> \frac{u_2(r_t r_p \underline{w}, 1 - \underline{w})}{u_1(r_t r_p \underline{w}, 1 - \underline{w})} \end{aligned}$$

Similarly, the second condition requires:

$$r_t r_p < \frac{u_2(r_t r_p \overline{w}, 1 - \overline{w})}{u_1(r_t r_p \overline{w}, 1 - \overline{w})}$$

Combining gives the result.

Proof of Proposition 3 Recall that $\hat{p}_J$ is a solution to:

$$G(p_J; c_J, w_J) = \tilde{c}(w_J, p_J) - c_J = 0 \tag{10}$$

and

$$\tilde{p}_J = \hat{p}_J(c(w_J, p_J), w_J).$$

So:

$$\begin{aligned}
\frac{\partial \tilde{p}_J}{\partial w_J} &= \frac{\partial \hat{p}}{\partial w_J} + \frac{\partial \hat{p}}{\partial c}\frac{\partial c}{\partial w_J} \\
&= -\frac{\frac{\partial G}{\partial w_J}}{\frac{\partial G}{\partial p_J}} + -\frac{\frac{\partial G}{\partial c_J}}{\frac{\partial G}{\partial p_J}}\frac{\partial c}{\partial w_J} \\
&= -\frac{\frac{\partial \tilde{c}}{\partial w_J}}{\frac{\partial \tilde{c}}{\partial p_J}} + \frac{1}{\frac{\partial \tilde{c}}{\partial p_J}}\frac{\partial c}{\partial w_J} \\
&= \frac{\frac{\partial c}{\partial w_J} - \frac{\partial \tilde{c}}{\partial w_J}}{\frac{\partial \tilde{c}}{\partial p_J}}
\end{aligned}$$

The numerator is strictly positive, and so the sign of the derivative is equal to the sign of the numerator. ∎

Proof of Proposition 4. If $w_A^{\text{br}}(r_t, \tilde{r}_p(\underline{w}+\epsilon)) = \underline{w}$ for some $\epsilon > 0$ or $w_A^{\text{br}}(r_t, \tilde{r}_p(\overline{w}-\epsilon)) = \overline{w}$ for some $\epsilon > 0$, then there is a stable corner equilibrium allocation. To complete the proof we need to show that if neither of these hold, there is an interior equilibrium. Let

$$F(w_A) = w_A^{\text{br}}(r_t, \tilde{r}_p(w_A)) - w_A \tag{11}$$

That is, $F(w_A)$ represents how he would change his allocation if starting from w_A, and an equilibrium is a point where $F(w_A^*) = 0$. If there is no stable corner solution, then it must be the case that $w_A^{\text{br}}(r_t, \tilde{r}_p(\underline{w}+\underline{\epsilon})) > \underline{w}$ for some small $\underline{\epsilon} \in (0, 1/2)$, and hence $F(\underline{w}+\underline{\epsilon}) > 0$. There must also be a $\overline{\epsilon} \in (0, 1/2)$ such that and $w_A^{\text{br}}(r_t, \tilde{r}_p(\overline{w}-\overline{\epsilon})) > 0$ and similarly $F(\overline{w}-\overline{\epsilon}) < 0$. By the continuity of w_A^{br} in $\tilde{r}_p$ and the continuity of $\tilde{r}_p$ in w_A, F is continuous in w_A, and so the intermediate value theorem implies there must be a $w_A^* \in (\underline{\epsilon}, \overline{\epsilon})$ such that $F(w_A^*) = 0$, where $F'(w_A) < 0$. Finally, since $F'(w_A) = \frac{\partial w_A^{\text{br}}}{\partial w_A} - 1$, then $F'(w_A^*) < 0 \iff \frac{\partial w_A^{\text{br}}}{\partial w_A}\Big|_{w_A = w_A^*} < 1$, and w_A^* is stable. ∎

In the main text, we describe the following result. Here, we state and prove it formally.

Proposition 8. *If the officer utility is given by equation 2, then there is a unique*

equilibrium in which the officer chooses a policing allocation

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

where

$$\widehat{w}_A = w_A^\dagger + \frac{\nu(r_t r_p - 1)}{(1-\nu)(1 + r_t r_p)} \tag{12}$$

and forms a (potentially inaccurate) belief $\tilde{r}_p^*$ using (4).

Proof of Proposition 8 Using Definition 1,an equilibrium policing allocation w_A solves

$$w_A^* = w_A^{\text{br}}(r_t, \tilde{r}_p^*(w_A^*))$$

At any interior solution, $w_A^{\text{br}}(r_t, r_p) = \frac{r_t \tilde{r}_p(w_A)}{1 + r_t \tilde{r}_p(w_A)}$. Substituting (4) and solving this equation for w_A gives a unique solution $\widehat{w}_A$, defined by equation (12) in the main text. Thus when $\widehat{w}_A$ lies in $[\underline{w}, \overline{w}]$ it meets the condition for a unique equilibrium allocation, $w_A^* = \widehat{w}_A$.

To prove that the corner solutions lie where the proposition claims, it helps to first describe the shape of the function which in turn describes how the allocation would change if playing an unconstrained best response starting at w_A,

$$F(w_A) = \frac{r_t \tilde{r}_p(w_A)}{1 + r_t \tilde{r}_p(w_A)} - w_A,$$

on the full range of $[0, 1]$. This function is continuous and differentiable. It is immediate that $F(0) = 0$ and $F(1) = 0$,[20] and by the analysis above $F(\widehat{w}_A) = 0$. So, when $\widehat{w}_A \in (0, 1)$, there are three zeroes on $[0, 1]$, and when $\widehat{w}_A$ lies outside of this interval the only zeroes are at the endpoints (and hence the function must be

[20]This implies that if we did not restrict the range to $[\underline{w}, \overline{w}]$, there would always be an equilibrium only policing either group, though this would not meet the stability condition whenever an interior equilibrium exists.

always positive or negative). Recall that:

$$\widehat{w}_A = \frac{r_t r_p}{1 + r_t r_p} + \frac{\nu(r_t r_p - 1)}{(1-\nu)(1 + r_t r_p)}$$

Rearranging and simplifying gives:

$$0 < \widehat{w}_A < 1 \iff \nu < r_t r_p < 1/\nu$$

In order to see whether F is positive or negative as $w_A \to 0$ and $w_A \to 1$, we need to check F' at these two points. Taking the first derivative of F yields:

$$F'(w_A) = \frac{\nu r_p r_t \left(2(1-\nu)w_A^2 - 2(1-\nu)w_A + 1\right)}{\left(\nu w_A^2 \left(r_p r_t + 1\right) - 2\nu w_A + \nu + (1 - w_A) w_A \left(r_p r_t + 1\right)\right)^2} - 1$$

Evaluating at 0 and 1 gives:

$$F'(0) > 0 \iff r_t r_p > \nu \qquad\qquad F'(1) > 0 \iff r_p r_t < \frac{1}{\nu}$$

Since $\nu < 1/\nu$, there are three cases we must consider, corresponding to three possible shapes of the F function. In case (I), $r_t r_p \geq 1/\nu$. When the inequality is strict, this implies F is increasing at 0, decreasing at 1, and has no interior root, and hence $F(w_A) > 0$ for $w_A \in (0, 1)$. When $r_t r_p = 1/\nu$, the only difference is that $F'(1) = 0$, but F is decreasing for w_A approaching 1, and this does not affect the rest of the argument. In case (II), $\nu < r_t r_p^2 < 1/\nu$, and so F is increasing at 0 and at 1, with an interior zero at $\widehat{w}_A$, and hence $F(w_A) > 0$ for $w_A \in (0, \widehat{w}_A)$ and $F(w_A) < 1$ for $w_A \in (\widehat{w}_A, 1)$. In case (III) $r_t r_p \leq \nu$, and F is decreasing at 0 (or, in the case where $r_t r_p = \nu$, flat at 0 but decreasing for small w_A), increasing at 1, and has no interior root, and hence $F(w_A) < 0$ for $w_A \in (0, 1)$. Note that there can only be an interior equilibrium in case (II), and it must be the case that $F'(\widehat{w}_A) < 0$, which is equivalent to, the stability condition.

Now we can complete proving where the equilibrium lies and uniqueness when the domain of the allocation choice is restricted to $[\underline{w}, \overline{w}]$. If $\widehat{w}_A \leq 0$ then the F function is in case (III) above, and so $F(\underline{w}) < 0$. If $0 < \widehat{w}_A < \underline{w}$, it is in case (II), but since $\underline{w} \in (\widehat{w}_A, 1)$ it must also be the case that $F(w_A) < 0$ for all $w_A \in [\underline{w}, \overline{w}]$.

And returning to the definition of w_A^{br}, $F(\underline{w}) < 0$, implies $w_A^{\text{br}}(r_t, \tilde{r}_p(\underline{w})) = \underline{w}$, meaning there is an extreme equilibrium at $\underline{w}$. $F(w_A) < 0$ also implies there is no interior equilibrium or equilibrium at $\overline{w}$ since $F(\overline{w}) < 0$, so this equilibrium is unique. If $\widehat{w}_A = \underline{w}$, then it is immediate that $F(\underline{w}) = \underline{w}$, and hence there is an extreme equilibrium at this bound, and this equilibrium is unique since $F(w_A) < 0$ for $w_A \in (\underline{w}, \overline{w}]$.

When $\underline{w} < \widehat{w}_A < \overline{w}$, $\widehat{w}_A$ is an interior equilibrium, and there can't be another interior state since there is no other point on $[\underline{w}, \overline{w}]$ where $F(w_A) = 0$. The F function is in case (II), which implies $F(\underline{w}) > 0$ and $F(\overline{w}) < 0$, so there is no equilibrium at the extremes. Thus the equilibrium is unique.

By a similar argument to the $\widehat{w}_A \leq \underline{w}$ case, if $\widehat{w}_A \geq \overline{w}$, then $w_A^{\text{br}}(r_t, \tilde{r}_p(\overline{w})) = \overline{w}$, and there can't be an equilibrium at $\underline{w}$ or on the interior. ∎

Proof of Proposition 5. Let $\nu \in (0, 1)$.

Part (i) immediately follows from the facts that $w_A^{\text{br}}(1, 1) = 1$ and $\tilde{r}_p(1/2) = 1$.

For part (ii) as in the proof of Proposition 4 let $F(w_A)$ be the difference between w_A and the best response allocation give the belief generated by w_A. If the equilibrium is unique, it must be stable by Proposition 4. And so if w_A^* is the equilibrium, it must be the case that $F(w_A) > 0$ if and only if $w_A < w_A^*$ and $F(w_A) < 0$ if and only if $w_A > w_A^*$.

From Lemma 1, if $r_t r_p < 1$ then $w_A^\dagger < 1/2$, and so $\tilde{r}_p(w_A^\dagger) < r_p$, and so $w_A^{\text{br}}(r_t, r_p) > w_A^{\text{br}}(r_t, \tilde{r}_p(w_A^\dagger))$, and $F(w_A^\dagger) < 0$. Therefore $w_A^\dagger > w_A^*$, and $\tilde{r}_p(w_A^*) < r_p$. The proof for $r_t r_p > 1$ follows an identical logic. ∎

Proof of Proposition 6. Write the equilibrium condition as:

$$F(w_A^*; r_t, \nu) = w_A^{\text{br}}(r_t, \tilde{r}_p(w_A^*)) - w_A^* = 0 \tag{13}$$

where we explicitly write F to be a function of exogenous parameters of interest (here, r_t and ν). Implicitly differentiating 13 with respect to ν gives:

$$\frac{\partial w_A^*}{\partial \nu} = \frac{\frac{\partial w_A^{\text{br}}}{\partial \tilde{r}_p} \frac{\partial \tilde{r}_p}{\partial \nu}}{1 - \frac{\partial w_A^{\text{br}}}{\partial \tilde{r}_p} \frac{\partial \tilde{r}_p}{\partial w_A^*}}$$

By the logic from above, the denominator of the right hand side is strictly positive. Then, this condition shows that $\frac{\partial w_A^*}{\partial \nu} > 0$ if and only if $\frac{\partial \tilde{r}_p}{\partial \nu} > 0$. However, from equation (4) in the main text, there are three cases: (1) $\frac{\partial \tilde{r}_p}{\partial \nu} > 0$ if $w_A^* > 1/2$, (2) $\frac{\partial \tilde{r}_p}{\partial \nu} < 0$ if $w_A^* < 1/2$ and (3) $\frac{\partial \tilde{r}_p}{\partial \nu} = 0$ if $w_A^* = 1/2$. Moreover, note that that if $\nu = 0$, then $w_A^* = w_A^\dagger$. Combining these observations: (1) if $w_A^* > 1/2$, then $|w_A^* - w_A^\dagger| = w_A^* - w_A^\dagger$ is increasing in ν since w_A^* is increasing away from $w_A^\dagger$ as ν increases, and (2) if $w_A^* < 1/2$, then $|w_A^* - w_A^\dagger| = w_A^\dagger - w_A^*$ is increasing in ν since w_A^* is decreasing away from $w_A^\dagger$ as ν increases. Finally, since (3) $w_A^* = w_A^\dagger$ for all ν if w_A^*, then we have shown that Δ^* is strictly increasing in ν if and only if $w_A^* \neq 1/2$.

∎

Proof of Proposition 7. To prove the existence of an equilibrium allocation, define a function $G : [\underline{w}, \overline{w}]^2 \to [\underline{w}, \overline{w}]^2$ given by

$$G(w_{A,1}, w_{A,2}) \equiv (w_A^{\text{br}}(r_{t,1}, \tilde{r}_{p,1}(w_{A,1}, w_{A,1}), w_A^{\text{br}}(r_{t,2}, \tilde{r}_{p,2}(w_{A,1}, w_{A,1})).$$

This is a continuous mapping from a compact and convex set to itself, so by the Brouwer fixed point theorem there must be a $(w_{A,1}^*, w_{A,2}^*)$, such that $G(w_{A,1}^*, w_{A,2}^*) = (w_{A,1}^*, w_{A,2}^*)$, which is an equilibrium allocation, with corresponding equilibrium beliefs given by $\tilde{r}_{p,i}^* = \tilde{r}_{p,i}(w_{A,1}^*, w_{A,2}^*)$.

We now show the comparative static results. First, recall we can write the equilibrium conditions as the following system of equations:

$$F_1(w_{A,1}, w_{A,2}; r_{t,1}, \nu_1) = w_A^{\text{br}}(r_{t,1}, \tilde{r}_{p,1}(w_{A,1}, w_{A,2})) - w_{A,1} = 0$$
$$F_2(w_{A,1}, w_{A,2}; r_{t,1}, \nu_1) = w_A^{\text{br}}(r_{t,2}, \tilde{r}_{p,2}(w_{A,1}, w_{A,2})) - w_{A,2} = 0$$

For part (i), we prove the result as $r_{t,1}$ changes, but identical logic holds for $r_{t,2}$.

To implicitly differentiate the equilibrium conditions with respect to $r_{t,1}$, take the total derivative of F_1 and F_2 (at w_A^*, accounting for the fact that $w_{A,i}$ are a

function of $r_{t,1}$:

$$\left.\frac{dF_1}{dr_{t,1}}\right|_{w_A=w_A^*} = \left(\left.\frac{\partial w_A^{\text{br}}}{\partial r_{t,1}}\right|_{w_A=w_A^*} + Y_1\left(Z_1\left.\frac{\partial w_{A,1}}{\partial r_{t,1}}\right|_{w_A=w_A^*} + Z_1\left.\frac{\partial w_{A,2}}{\partial r_{t,1}}\right|_{w_A=w_A^*}\right)\right) - \left.\frac{\partial w_{A,1}}{\partial r_{t,1}}\right|_{w_A=w_A^*} = 0 \tag{14}$$

$$\left.\frac{dF_2}{dr_{t,1}}\right|_{w_A=w_A^*} = \left(Y_1\left(Z_2\left.\frac{\partial w_{A,1}}{\partial r_{t,1}}\right|_{w_A=w_A^*} + Z_2\left.\frac{\partial w_{A,2}}{\partial r_{t,1}}\right|_{w_A=w_A^*}\right)\right) - \left.\frac{\partial w_{A,2}}{\partial r_{t,1}}\right|_{w_A=w_A^*} = 0 \tag{15}$$

where as in section B we define:

$$Y_i = \left.\frac{\partial w_A^{\text{br}}}{\partial \tilde{r}_{p,i}}\right|_{\tilde{r}_{p,i}=\tilde{r}_{p,i}(w_{A,1}^*,w_{A,2}^*)}$$
$$Z_i = \left.\frac{\partial \tilde{r}_{p,i}(w_{A,1},w_{A,2})}{\partial w_{A,1}}\right|_{w_A=w_A^*} = \left.\frac{\partial \tilde{r}_{p,i}(w_{A,1},w_{A,2})}{\partial w_{A,2}}\right|_{w_A=w_A^*}.$$

Equations (14) and (15) are a system of two equations where we want to solve for $\frac{\partial w_{A,1}}{\partial r_{t,1}}$ and $\frac{\partial w_{A,1}}{\partial r_{t,1}}$. Define the following:

$$T_1 = \left.\frac{\partial w_{A,1}}{\partial r_{t,1}}\right|_{w_A=w_A^*} \qquad T_2 = \left.\frac{\partial w_{A,2}}{\partial r_{t,1}}\right|_{w_A=w_A^*} \qquad X = \left.\frac{\partial w_A^{\text{br}}}{\partial r_{t,1}}\right|_{w_A=w_A^*}$$

Then, we can rewrite this system of equations as

$$(X + Y_1Z_1(T_1 + T_2)) - T_1 = 0$$
$$Y_1Z_1(T_1 + T_2) - T_1 = 0$$

and goal is to solve for T_1 and T_2. This gives:

$$T_1 = X + \frac{XY_1Z_1}{1 - Y_1Z_1 - Y_2Z_2}$$
$$T_2 = \frac{XY_2Z_2}{1 - Y_1Z_1 - Y_2Z_2}.$$

Since we know that $X > 0$, $Y_i > 0$, and $Z_i > 0$, both of these are strictly positive

if and only if $1 - Y_1 Z_1 - Y_2 Z_2 > 0$, which is exactly the stability condition for an interior equilibrium derived in section B. Finally, since $\Delta^* = |w^*_{A,i} - w^\dagger_{A,i}|$ and $w^\dagger_{A,i}$ is constant in $r_{t,1}$, then for each $i \in \{1, 2\}$, Δ^* increases in $r_{t,1}$.

For part (ii), we prove the result as ν_1 changes, but identical logic holds for ν_2. We now define the following:

$$N_1 = \left.\frac{\partial w_{A,1}}{\partial \nu_1}\right|_{w_A = w^*_A} \qquad\qquad N_2 = \left.\frac{\partial w_{A,2}}{\partial \nu_1}\right|_{w_A = w^*_A}$$

To implicitly differentiate the equilibrium conditions with respect to ν_1, take the total derivative of the equilibrium conditions at w^*_A, accounting for the fact that $w_{A,i}$ is a function of ν_1:

$$Y_1\left(Z_1 N_1 + Z_1 N_2 + \frac{\partial \tilde{r}_{p,1}}{\partial \nu_1}\right) - N_1 = 0 \qquad Y_2\left(Z_2 N_1 + Z_2 N_2\right) - N_1 = 0$$

Our goal is to solve for N_1 and N_2, which gives:

$$N_1 = \frac{\partial \tilde{r}_{p,1}}{\partial \nu_1}\left(\frac{Y_1(1 - Y_2 Z_2)}{1 - Y_1 Z_1 - Y_2 Z_2}\right) \qquad N_2 = \frac{\partial \tilde{r}_{p,1}}{\partial \nu_1}\left(\frac{Y_1 Y_2 Z_2}{1 - Y_1 Z_1 - Y_2 Z_2}\right)$$

Again since we know that $X > 0$, $Y_i > 0$, and $Z_i > 0$, both of these are strictly positive at an interior equilibrium if and only if the stability condition is met and $\frac{\partial \tilde{r}_{p,1}}{\partial \nu_1} > 0$. This latter condition holds if $w_A = w_{A,1} + w_{A,2} > 1$ (i.e., group A receives a higher allocation than group B). Similarly, if $w_A < 1$, then $\frac{\partial \tilde{r}_{p,1}}{\partial \nu_1} < 0$ and hence both officers police group B more as ν_1 increases. ∎

D More General Beliefs (Multiple Officer Model)

There are several ways one could extend the definiton of non-conditioning bias to the multiple officer model. One potentially realistic change would be to assume that officers may do a better (or worse) job of adjusting for their own behavior than others' behavior when forming inferences about the p_J parameters. Formally, we

could define the officer belief as:

$$\tilde{r}_{p,i}(w_A) = \frac{\frac{c_A}{\nu_i^s + (1-\nu_i^s) w_{A,i} + \nu_i^o + (1-\nu_i^o) w_{j,2}}}{\frac{c_B}{\nu_i^s + (1-\nu_i^s) w_{B,i} + \nu_i^o + (1-\nu_i^o) w_{B,j}}} \tag{16}$$

where the $\nu_i^s \in [0,1]$ represents how well the officer conditions for his own allocation and $\nu_i^o \in [0,1]$ represents how well he conditions on the other officer choice. A key feature of this more general belief is that as long as $\nu_i^s > 0$, it is increasing in $w_{A,i}$, meaning the officer's belief about A's relative crime rate increases in how much he polices this group. Similarly, as long as $\nu_i^o > 0$, the officer's belief about the relative crime rate of group A increases in how much the other officer polices this group. So, while the the analysis is more complicated with this belief formation, the general feedback loop and spillover dynamics are present here as well.

E Nonlinear Returns to Policing

Returning to the original utility function, recall an additional way to motivate the diminishing returns assumption is that the marginal rate of crimes caught among group J decreases as w_J increases. Suppose the number of crimes caught is equal to $c_J = f(p_J w_J)$ where f is an increasing and concave function. Assume that the officers knows this functional form, but not the p_J parameters.

Knowing c_J and w_J, a fully Bayesian officer could then infer p_J by inverting the f function: $p_J = f^{-1}(c_J)/w_J$. The officer would then form a correct inference about the relative crime "rates" of the group, where the scare quotes highlight that the p_J parameters no long have a simple interpretation as the average crime rates of the groups:

$$\tilde{r}_p(0) = \frac{f^{-1}(c_A)/w_A}{f^{-1}(c_B)/w_B} = p_A/p_B$$

Note that if the officer now beliefs the relative crime rates are equal to c_A/c_B, he is making two mistakes: not adjusting for w_J, and also not accounting for the nonlinear effect of policing effort. In this case his belief about the relative prevalence

of crime among members of each group (as a function of the allocation decision) becomes:

$$\frac{f(p_A w_A)}{f(p_B(w - w_A))}$$

Which, as long as f is increasing, is increasing in w_A. Unfortunately with this notion of naivety there is not a natural way to come up with an "intermediate" form of the bias.

One potentially instructive special case is if f is a power function: $f(p_A w_A) = (p_A w_A)^\alpha$, $\alpha \in (0, 1)$. In this case the fully naive belief simplifies to:

$$\frac{(p_A w_A)^\alpha}{(p_B(w - w_A))^\alpha} = r_p^\alpha \left(\frac{w}{w - w_A}\right)^\alpha$$

If $r_p = 1$ this belief will be correct when $w_A = 1/2$ (and, so with no animus, the officer will again pick a correct allocation). Now when $r_p > 1$, $1 < r_p^\alpha < r_p$. So, if the officer were to allocate his time evenly between the groups, he would now *underestimate* the relative prevalence of crime among members of the group with the higher crime rate. In other words, "not understanding diminishing returns" could lead to the opposite effect as the bias we study.

Another way to model a naive officer is that he is able to "invert" the f function but does not account for the differential policing rate. Such an officer's belief becomes:

$$\frac{p_A w_A}{p_B(w - w_A)}$$

as in the baseline, so we can again define the intermediate form of naivety identically.

References

Gintis, Herbert. 2009. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior.* 2nd ed. Princeton University Press.

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.