A Causal Framework To Evaluate Racial Bias in Law Enforcement Systems

Jessy Xinyi Han[1], Andrew Cesare Miller[2], S. Craig Watkins[3], Christopher Winship[4], Fotini Christia[1], Devavrat Shah[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]United States Naval Academy, Annapolis, MD, USA
[3]University of Texas at Austin, Austin, TX, USA
[4]Harvard University, Cambridge, MA, USA

Abstract

We are interested in developing a data-driven method to evaluate race-induced biases in law enforcement systems. While the recent works have addressed this question in the context of police-civilian interactions using *police stop* data, they have two key limitations. First, bias can only be properly quantified if true *criminality* is accounted for in addition to race, but it is absent in prior works. Second, law enforcement systems are multi-stage and hence it is important to isolate the true source of bias within the "causal chain of interactions" rather than simply focusing on the end outcome; this can help guide reforms.

In this work, we address these challenges by presenting a multi-stage causal framework incorporating criminality. We provide a theoretical characterization and an associated data-driven method to evaluate (a) the presence of any form of racial bias, and (b) if so, the primary source of such a bias in terms of race and criminality. Our framework identifies three canonical scenarios with distinct characteristics: in settings like (1) airport security, the primary source of observed bias against a race is likely to be bias in law enforcement against *innocents* of that race; (2) AI-empowered policing, the primary source of observed bias against a race is likely to be bias in law enforcement against *criminals* of that race; and (3) police-civilian interaction, the primary source of observed bias against a race could be bias in law enforcement against that race or bias from the general public in reporting (e.g. via 911 calls) against *the other* race. Through an extensive empirical study using police-civilian interaction (stop) data and 911 call data, we find an instance of such a counter-intuitive phenomenon: in New Orleans, the observed bias is *against* the majority race and the likely reason for it is the over-reporting (via 911 calls) of incidents involving the minority race by the general public.

1 Introduction

The role played by race, and more generally, gender, religion, and ethnicity, in decision-making within societal systems has garnered renewed interests over the past decade. Among various societal systems, law enforcement plays a crucial role in maintaining public safety and order but is, like any other, susceptible to systemic bias (Skogan and Frydl 2004). Racial disparities in law enforcement are believed to perpetuate cycles of poverty, exclusion, and marginalization of minority communities. Therefore, a thorough examination of racial disparities and guidance for reforms to reduce, or ideally eliminate, such disparities is required (Goff and Kahn 2012; Hellman 2020; White 2019).

Towards this, the research community has been focusing on benchmarking data, understanding the underlying causes of biases and proposing policy changes to mitigate them (Horel et al. 2021; Goncalves and Mello 2021; Blair et al. 2021). In particular, racial disparities in law enforcement are often associated with a higher level of law enforcement actions against the discriminated race (Gelman, Fagan, and Kiss 2007; Ridgeway et al. 2007). Indeed, such an outcome specific inspection led Fryer (2019) to conclude a lack of racial bias against minorities on lethal use of force. This directly contradicts the public perception of racial disparities. To correct for this, subsequent works like Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020) have pointed out the need for a causal framework to correct for bias in such analyses and proposed appropriate causal estimands to evaluate the presence of racial bias or disparity, which are important steps in the right direction. However, these prior works have two key limitations.

First, they do not explicitly account for the notion of *criminality*, i.e. whether a crime was committed. To evaluate racial disparity, it is important to identify whether the use of force was justified, which requires the context whether a crime was committed. Indeed, in our opinion, the absence of criminality in the prior works was the primary reason for the public debate to arise around the identification of causal estimand to evaluate racial bias or disparity (see Gelman (2020a) and Gelman (2020b)).

Second, law enforcement is a multi-stage system. While the use of force by police is the eventual measurement to determine racial disparity, the incidents in which the police interact with civilians are determined by a sequence of decisions. This sequence includes which incidents are reported through 911 calls, how police are dispatched, and whether police stop civilians involved and eventually exert force against them. While Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020) have introduced one of the stages of this multi-stage system compared to Fryer (2019), further stages need to be introduced.

Contributions. In this work, we address these two challenges by presenting a multi-stage causal framework incorporating criminality. While criminality often remains unobserved or unverifiable, its inclusion in our framework enables us to define the appropriate causal estimand associated with racial disparity or bias. The proof of the pudding is, however, in the eating: the need to use *observed* data to evaluate the presence of racial disparity.

Towards that, we posit the following natural hypothesis (or assumption): given the *context* which may represent social, economic and all other conditions that might lead an individual to commit a crime, the race of an individual and the act of committing a crime are (conditionally) independent. In other words, the reasons behind someone committing a crime are determined by the surrounding context, not by their race.

Under this hypothesis, despite the notion of criminality being unobserved, we are able to define a test statistic of *observational racial disparity* (see (6) for a precise definition) using observed data only. The *observational racial disparity* evaluates whether there is racial disparity but falls short in identifying the primary cause of the disparity - which race is discriminated against, whether the criminals or innocents, and who is primarily responsible for such a discrimination. Indeed, in general such identification is beyond reach. However, we identify three canonical scenarios under which the primary cause of racial disparity can be discerned:

(1) Scenario 1 corresponds to settings like airport security checks, where everyone entering the system undergoes a check (i.e. interaction with law enforcement). In this scenario, if there is observational racial disparity based on the test statistic, it is primarily due to bias in law enforcement actions against innocent individuals of that race. See Section 3.2 for details.

(2) Scenario 2 corresponds to settings like AI-empowered policing where an individual is suspected and put through further interaction with law enforcement only if there is a high chance of requiring further investigation. In this scenario, if there is observational racial disparity based on the test statistic, then it is primarily due to the bias in law enforcement actions against criminal individuals of that race. See Section 3.3 for details.

(3) Scenario 3 corresponds to settings like police-civilian interactions where biases can arise from public reporting through 911 calls and/or law enforcement by police. In this scenario, we find that observed racial bias against one race may come from policing actions against that race or the general public's over-reporting of the other race. See Section 3.4 for details.

We apply our framework to the law enforcement system with police-civilian interaction using the police stop data and 911 call data for New York City (NYC) and New Orleans over a number of years. Through careful processing and stitching such data, we construct the multi-stage dataset of various incidences, corresponding to Scenario 3 shown above. Upon careful evaluation of the *observational racial disparity* for both NYC and New Orleans, we find the following:

In NYC, there are observational biases against minority races (Black/Hispanic) compared to the majority (White), which could be explained by the racial disparity in policing actions. However, in New Orleans we find the exact opposite - an observational bias against the majority race as per the observed test statistic. This could be explained by the over-reporting of incidents involving minorities through 911 calls, as suggested by our framework under Scenario 3. This leads us to a counter-intuitive phenomenon - even though at its core there is bias against the minority (in reporting), the observational data suggests a bias against the majority!

Related Prior Works. A growing literature with causal inference has been developed to model multi-stage processes in sociology (Elwert and Winship 2014), healthcare (Huang and Ning 2012), and political science (Slough 2022). Among them, an active area of research is the investigation of racial bias in law enforcement processes. Various methods have been been proposed (Neil and Winship 2019), including the benchmark test (Gelman, Fagan, and Kiss 2007), outcome test (Fryer 2019), threshold test (Simoiu, Corbett-Davies, and Goel 2017) and so on.

Most previous works evaluating potential bias in law enforcement systems focus on individual stages of the process, including traffic stops (Pierson et al. 2020; Antonovics and Knight 2009; Grogger and Ridgeway 2006), vehicle searches (Knowles, Persico, and Todd 2001; Ridgeway 2006), verbal communication (Voigt et al. 2017), arrests (Mitchell and Caudy 2015), sentencing (Alesina and La Ferrara 2014), and use of force (Nix et al. 2017; Kramer and Remster 2018), among other outcomes. Some recent studies do take a systematic approach to evaluate racial disparities. Specifically, Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020) utilize a causal framework to refine the simplistic approach in Fryer (2019) to evaluate racial disparity based on the use of force by police in police-civilian interactions with the police stop data. However, these works have two limitations as mentioned earlier: they do not account for the notion of *criminality* and fall short of encompassing some stages of law enforcement, including how incidents are reported. Consequently, they might suggest a bias against the majority, even though the actual source of bias could be the over-reporting of incidents involving minorities.

When comparing our causal framework and the observational racial disparity statistics with Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020), several key differences emerge. First, the causal frameworks in Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020) are subsets of our causal framework as they do not account for criminality and other stages of causal framework. Second, when the framework is restricted to the simpler setting in Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020), our test statistics induces similar test statistics suggested in their works. However, Gaebler et al. (2020) requires an additional ignorability assumption to justify the identification of their statistics while our work naturally derives such justification from the hypothesis that race and criminality of an individual are conditionally independent given the context. Like Gaebler et al. (2020), Knox, Lowe, and Mummolo (2020) needs to make additional assumptions to justify their data-driven approach for point-identification of their causal

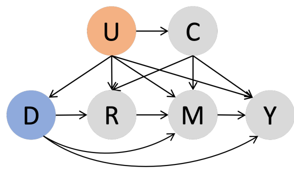


Figure 1: A causal DAG correpsonding to multi-stage law enforcement system.

Notation	Meaning	Observed?
D	race	yes
R	report	yes
M	stop	yes
Y	policing outcome	yes
U	context	partially yes
C	criminality	no

Table 1: A summary of notations.

estimand, unlike our work. Our framework identifies three canonical scenarios where observational racial disparity can be explained through the primary source of racial disparity across stages of law enforcement and the types of disparity (against race and criminals / innocents).

Since differential crime reporting rates have significant impacts on policing systems (Akpinar, De-Arteaga, and Chouldechova 2021), another salient feature of our work is the ability to leverage the multi-stage nature of law enforcement by including public reporting through 911 call data. Some existing studies have leveraged the reporting party to assess possible discrimination but they focus on 911 calls themselves and are primarily concerned with teasing out biases based on officer-level characteristics. For example, Hoekstra and Sloan (2022) uses 911 calls as a natural experiment to assess the effect of officer race on use of force likelihood. Weisburst (2022) leverages 911 calls to evaluate how "officer arrest propensity" influences police discretion. These works do not leverage—nor aim to do so—the reporting decisions to measure levels of overall racial disparities in law enforcement actions.

Organization. The remainder of the article is organized as follows: In Section 2, we lay out a systematic causal framework for evaluating bias in law enforcement. In Section 2.4, we explain how prior works fit within this framework. In Section 3, we discuss conditions under which we could identify the primary source of racial disparities in different law enforcement systems. In Section 4, we explain our empirical strategy, analyze our data and present the results. In Section 5, we conclude with a discussion on the potential extensions of our framework and reflect on the broader implications of our work.

2 A Causal Framework for Racial (Dis)Parity

Law enforcement systems are complex and often involve multiple stages. To comprehend the various intricacies of these systems, including the presence of racial disparities, we need to closely study the inter-relationship between these stages. In this section, we introduce a multi-stage causal framework.

2.1 Causal Framework

Consider a collection of incidents that demand the attention of law enforcement agencies. As mentioned earlier, law enforcement is a *multi-stage* process, involving not only the law enforcement agency but also the reporting party who may report only a portion of the incidents. Consequently, only the reported incidents will receive appropriate law enforcement attention and outcomes. For example, in airport security checks, individuals intending to board a plane self-report to Transportation Security Administration (TSA) officers and undergo mandatory security checks; in advanced AI-empowered policing, AI algorithms detect traces of criminal activities and report them to the police who later enforce policing actions; in a typical police-civilian encounter, as we find using empirical analysis (see Section 4 for details), that the majority of incidents that come to law enforcement's attention originate from 911 calls. Once the interaction is initiated in any of these scenarios, the law enforcement agents decide whether to stop and further question the suspects and, depending on the interaction, whether to proceed with potentially severe actions. We formalize this multi-stage law-enforcement process through a natural causal framework represented by a causal Directed Acyclic Graph (DAG) in Figure 1 with pertinent notations summarized in Table 1.

Incident. An incident involves a person or people of interest, and we assume each incident involves a single person of interest. Let $D \in \{\text{majority}, \text{minority}\}$ denote the race (or type) of the individual associated with the incident. We specifically analyze the effect of race in this work.[1]

Reporting. An individual or incident is reported to the law enforcement system. We use $R = 1$ to indicate an incident is reported and $R = 0$ otherwise.

Stop and Law Enforcement. Following the reporting of an incident to the law enforcement agency, the individual involved may be stopped for further inquiry. Let $M = 1$ denote the incident that leads to a stop and $M = 0$ otherwise. If the stop occurs ($M = 1$), we denote by $Y = 1$ the enforcement of some subsequent action (e.g., verbal warning, use of force, citation, arrest) and $Y = 0$ otherwise.

Criminality. Law enforcement systems aim to hold individuals accountable for crimes committed. Let $C \in \{0, 1\}$ denote whether the individual commits the crime ($C = 1$) or not ($C = 0$).

Context. The context of each incident, such as the socio-economic characteristics and location, that may influence the involved civilian's race, reporting, stopping, and law enforcement actions. We denote by $U \in \mathbb{R}^l$ the

[1]Much scholarship on race has demonstrated the constructivist nature of racial categorizations (Sen and Wasow 2016). We are prevented from taking a constructivist approach given data limitations in large-scale law enforcement datasets. Moreover, causally-oriented literature on biases in law enforcement also generally takes a non-constructivist approach, so we follow suit in order to speak to that literature.

$l-$dimensional context that impacts the race D, criminality C, reporting R, stop M, and law-enforcement action Y.

Causal Mechanism. As noted, U may influence D, C, R, M, and Y. The true nature of criminality C obviously could affect the reporting R, stop M, and law-enforcement action Y. Reporting R influences the stop M, and through M, it influences the law-enforcement action Y. The race D may affect the reporting R, stop M, and law-enforcement action Y. The underlying model is represented in a causal Directed Acyclic Graph (DAG) in Figure 1. The key (and only) assumption encoded in this causal mechanism is that there is *no edge* between C, D. That is, given *context* U, criminality C and race D are *conditionally independent*. In words, this encodes

it is the context that makes someone criminal (or not), not their race.

2.2 Achieving Racial Parity Across Stages

Given the causal DAG, we study the "ideal" racial parity to understand how the *intervention* of race D affects the reporting outcome R and law-enforcement outcome Y[2]. Specifically, adopting the standard causal notation (Pearl 2014), we define racial parity in reporting and in law enforcement.

Let $\mathbb{P}^{\mathrm{do}(D=d)}(\cdot)$ represent the distribution induced under the intervention of race $D = d$ for $d \in \{\text{majority}, \text{minority}\}$.

Definition 1 (Racial parity in reporting)**.** *There is* racial parity in reporting *with respect to majority and minority citizens if* $\forall u \in \mathbb{R}^l$ *and* $c \in \{0, 1\}$*,*

$$\mathbb{P}^{do(D=majority)}[R = 1 \mid C = c, U = u] = \mathbb{P}^{do(D=minority)}[R = 1 \mid C = c, U = u]. \tag{1}$$

Definition 2 (Racial parity in law enforcement)**.** *There is* racial parity in law enforcement *with respect to majority and minority citizens if* $\forall u \in \mathbb{R}^l$ *and* $c \in \{0, 1\}$*,*

$$\mathbb{P}^{do(D=majority)}[Y = 1 \mid R = 1, C = c, U = u] = \mathbb{P}^{do(D=minority)}[Y = 1 \mid R = 1, C = c, U = u]. \tag{2}$$

In this work, we aim to understand if either parity holds and, if not, whether there is bias against citizens from the majority or minority racial group with respect to criminals or innocents or both in either stage. Note that depending on the scenarios, there are many different definitions of fair treatment in literature (Verma and Rubin 2018; Kilbertus et al. 2017; Kusner et al. 2017). The notion of parity introduced in Definition 1 and 2 is a counterfactual version of *equalized odds* (Hardt et al. 2016) often used in criminology literature (Blumstein and Cohen 1979).

[2]The notion of manipulating a person's race is understandably controversial. A possible solution is to shift from immutable traits to perceptions of them (Greiner and Rubin 2011). Thus, in this work, we model a scenario where two individuals in the same context are identical except for the perception of their races and understand whether this leads to different law-enforcement actions.

As per the causal DAG in Fig. 1 and do calculus, we could represent the causal quantity in Definition 1 and 2 as the following:

Lemma 1. $\forall u \in \mathbb{R}^l$*,* $c \in \{0, 1\}$*, and* $d \in \{$*minority*, *majority*$\}$*,*

$$\mathbb{P}^{do(D=d)}[R = 1 \mid C = c, U = u] = \mathbb{P}[R = 1 \mid D = d, C = c, U = u]. \tag{3}$$

In words, the causal quantity in Definition 1 racial parity in reporting is equivalent to the probability of incidents being reported by the reporting party (e.g. 911 callers), given context $U = u$ and criminality $C = c$. This essentially captures the role of the reporting party in law enforcement systems.

Lemma 2. $\forall u \in \mathbb{R}^l$*,* $c \in \{0, 1\}$*, and* $d \in \{$*minority*, *majority*$\}$*,*

$$\begin{aligned} &\mathbb{P}^{do(D=d)}[Y = 1 \mid R = 1, C = c, U = u] \\ &= \mathbb{P}[Y = 1 \mid R = 1, D = d, C = c, U = u] \\ &= \mathbb{P}[Y = 1, M = 1 \mid R = 1, D = d, C = c, U = u]. \end{aligned} \tag{4}$$

In words, the causal quantity in Definition 2 racial parity in law enforcement is equivalent to the likelihood of an individual being stopped and subject to law-enforcement actions, given that the incident is reported and the context $U = u$ and criminality $C = c$ are controlled. As such, it primarily captures the role of law enforcement agents, such as police who are responsible for stops, use of force, and other related actions.

Though different notions of fairness can often be incompatible, given the causal DAG in Fig. 1, parity in public reporting leads to other desirable definitions of fairness like *predictive parity* and *group fairness* (Verma and Rubin 2018; Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2016). Details of the proof can be found in Appendix Section A.

The challenge with the above definitions is that criminality C is never fully known. Therefore, it may never be feasible to definitely confirm (or reject) parity in either law enforcement or reporting under this framework. Despite this limitation, next, we discuss certain invariances involving observed quantities that ought to be preserved. And, they can lead to test statistics for verifying racial parity.

2.3 Verifying Racial Parity With Observed Data

The existence of racial parity as defined in Definitions 1 and 2 are not verifiable using observed quantities or data. However, they do imply certain invariances involving observed quantities only. We shall call them observational parity.

Definition 3 (Observational parity in law enforcement)**.** *There is* observational racial parity in law enforcement *if*

$$\mathbb{P}[Y = 1, M = 1 \mid R = 1, D = \textit{majority}, U = u] = \mathbb{P}[Y = 1, M = 1 \mid R = 1, D = \textit{minority}, U = u], \;\; \forall u \in \mathbb{R}^l. \tag{5}$$

We similarly define *observational parity in law enforcement in context* u as the satisfaction of Eq.(5) for the given context u. This observational parity suggests that a reported majority person and a reported minority person are equally likely to be stopped and later subject to law enforcement actions. We point out that key distinction between Definition 2/1 and observational parity in Definition 3 is that the former consider criminality C while the latter does not. The nuance enables us to directly estimate the probabilities in Eq.(5) from observational data without requiring information on criminality. Towards that, we state the following result that connects racial parity (that can not be evaluated) with observational racial parity. The proof can be found in Appendix Section B.

Proposition 1. *If there is racial parity in both law enforcement and reporting, there is observational parity in law enforcement.*

The above result suggests a natural statistical test. Specifically, for each context $u \in \mathbb{R}^l$, the (5) needs to be satisfied if the racial parity holds. Therefore, violation of it for any context $u \in \mathbb{R}^l$ would imply the lack of racial parity. We formally state this next and its proof can be found in Appendix Section C.

Theorem 1. *If* $\exists u \in \mathbb{R}^l$ *such that the observational parity in law enforcement in context* u *is violated, i.e.* (5) *does not hold, then there cannot be parity in law enforcement and reporting simultaneously in the given context* u*.*

With parity in law enforcement and reporting, a reported individual should be equally likely to be stopped and later subject to law enforcement actions regardless of his/her race. In other words, if we find a lack of such observational parity, even *without observing* criminality C, the true parity considering criminality C in either reporting or law enforcement must be violated. Given this, we define the following *test statistics*.

Definition 4 (Observational racial disparity)**.** $\forall u \in \mathbb{R}^l$, *let* observational racial disparity*, denoted as* $\Delta(u)$*, be*

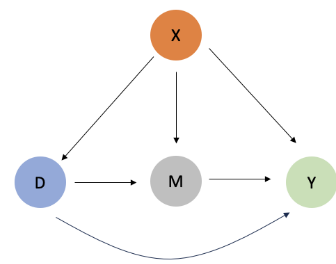
$$\Delta(u) := \mathbb{P}[Y = 1, M = 1 \mid R = 1, D = \textit{majority}, U = u] - \mathbb{P}[Y = 1, M = 1 \mid R = 1, D = \textit{minority}, U = u]. \quad (6)$$

In effect, $\Delta(u)$ captures the difference in the likelihood of facing a law enforcement action such as use of force between races majority and minority assuming the individual is reported and stopped.

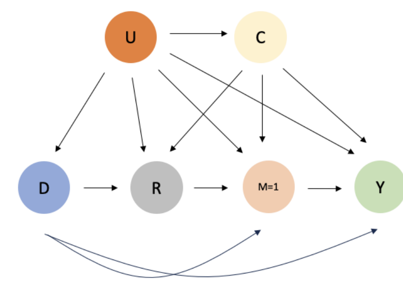
Note that we do not assume the composition of reported incidents $R = 1$ to be the same as the composition of unreported incidents $R = 0$. The purpose of evaluating (5) and (6), quantities conditioning on the post-treatment variable $R = 1$, is purely for inferring the presence or absence of parity in law enforcement and reporting, which is our real interest. This is vastly different from treating (5) as the ultimate fairness definition of interest, which is the problem motivating (Knox, Lowe, and Mummolo 2020).

2.4 Situating Prior Works in Our Framework

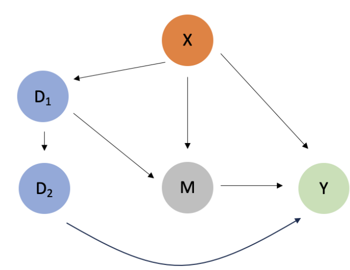
We discuss how prior works, specifically Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020), fit within our framework.



(a) Causal DAG corresponding to prior works (Knox, Lowe, and Mummolo 2020; Gaebler et al. 2020).



(b) Reduced form of Causal DAG of this work.



(c) Our interpretation of causal DAG utilized by (Knox, Lowe, and Mummolo 2020) to define counterfactual notion of racial parity.

Figure 2: Variations of DAGs to situate prior works in our framework.

Comparing Causal DAGs. As discussed earlier, these prior works have two limitations. First, they do not account for criminality explicitly. Second, accounting for multiple stages, specifically reporting, is absent. The causal DAG in these prior works can be abstracted as the one described in Figure 2a. The causal DAG in Figure 2b describes the *reduced* form of our mechanism where in DAG of Figure 1, we set $M = 1$ with probability 1. It can be seen that, with the exception of presence of C, the DAGs in Figure 2a and Figure 2b are identical with mapping $X \leftrightarrow U$ and $M \leftrightarrow R$.

Comparing Notions of Racial Parity. We recall the notions of racial parity in (Knox, Lowe, and Mummolo 2020; Gaebler et al. 2020) and compare them with the notion of racial parity of Section 2.2. In Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020), like in this work, racial parity is defined through a counterfactual notion induced through intervention on race.

In Gaebler et al. (2020), the racial parity is defined through causal estimand $\mathrm{CDE}_{\mathrm{M}=1}$ defined as

$$\begin{aligned}\mathrm{CDE}_{\mathrm{M}=1} &= \mathbb{P}^{do(D=\text{majority},M=1)}[Y=1|M=1,X=x] \\ &\quad - \mathbb{P}^{do(D=\text{minority},M=1)}[Y=1|M=1,X=x].\end{aligned} \tag{7}$$

Under Causal DAG of Figure 2a, per the somewhat controversial ignorability assumption $Y^{do(D=d,M=1)} \perp D|X, M=1$ made in Gaebler et al. (2020), it follows that

$$\begin{aligned}&\mathbb{P}^{do(D=d,M=1)}[Y=1|M=1,X=x] \\ &= \mathbb{P}[Y=1|D=d,M=1,X=x].\end{aligned} \tag{8}$$

That is,

$$\begin{aligned}\mathrm{CDE}_{\mathrm{M}=1} &= \mathbb{P}[Y=1|D=\text{majority},M=1,X=x] \\ &\quad - \mathbb{P}[Y=1|D=\text{minority},M=1,X=x].\end{aligned} \tag{9}$$

Under the mapping described above and explained through Figure 2a and Figure 2b, it can be checked that (9) is equivalent to our test statistics in (6) as defined in Definition 4.

Moreover, Knox, Lowe, and Mummolo (2020) considers another notion of racial parity defined through causal estimand $\mathrm{ATE}_{\mathrm{M}=1}$. To define it, authors introduce a variation of the counterfactual intervention on race.

While Knox, Lowe, and Mummolo (2020) do not make this explicit, to explain their approach we introduce a “modified” causal DAG as in Figure 2c which is obtained by splitting variable D in Figure 2a into two variables D_1, D_2 with edge between D_1 and D_2 encoding relationship $D_2 = D_1$. With respect to this DAG, the causal estimand $\mathrm{ATE}_{\mathrm{M}=1}$ can be defined as follows:

$$\begin{aligned}\mathrm{ATE}_{\mathrm{M}=1} &= \mathbb{P}^{do(D_2=\text{majority})}[Y=1|M=1,X=x] \\ &\quad - \mathbb{P}^{do(D_2=\text{minority})}[Y=1|M=1,X=x].\end{aligned} \tag{10}$$

While under the original causal DAG of Figure 2a, with $D = D_1 = D_2$, the above would have reduced to a similar quantity as in $\mathrm{CDE}_{\mathrm{M}=1}$ of (9), the modification makes it different. In particular, despite assuming it to be causal DAG, it can not be point-identified since we do not have two different D_1, D_2 in reality. Therefore, (Knox, Lowe, and Mummolo 2020) take a partial identification approach towards $\mathrm{ATE}_{\mathrm{M}=1}$ and derive nonparametric sharp bounds.

In Summary. The two important prior works Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020) can be explained through the reduced form of our framework. In the reduced form to match the prior work, our test statistics matches with that of Gaebler et al. (2020). In that sense, ours can be viewed as a natural generalization of Gaebler et al. (2020) to account for criminality as well as multi-stage effect. The causal estimand of Knox, Lowe, and Mummolo (2020) can also be explained within the same framework, however it requires additional assumptions and data to lead to point-identification using observed data.

3 Measuring Racial Disparity: Law Enforcement, Reporting

The Theorem 1 of Section 2 provides test statistics to verify whether there is racial parity or not. Specifically, if for any context $u \in \mathbb{R}^l$, if $\Delta(u)$ as defined in (6) is $\neq 0$, then racial disparity does exist. However, it does not help determine the primary source of disparity nor its intensity.

Indeed, in general, it is infeasible to pinpoint the primary source and intensity of disparity given that not everything is observed. However, in this section, we shall discuss three canonical scenarios where it would be feasible to determine the primary source of disparity and intensity of racial disparity.

To that end, we define the proportion of innocents ($C=0$) being reported ($R=1$) for a given context $u \in \mathbb{R}^l$ and race $d \in \{\text{minority}, \text{majority}\}$ as

$$\xi_{d,u} = \mathbb{P}[C=0 \mid R=1, D=d, U=u]. \tag{11}$$

This quantity suggests three canonical scenarios:

Scenario 1 (aka Airport security checks): $\forall d \in \{\textit{majority}, \textit{minority}\}, \forall u \in \mathbb{R}^l, \xi_{d,u} \simeq 1$, i.e., most reported individuals are innocent. This naturally happens in settings such as airport security check with everyone coming to the airport voluntarily submits to security check (i.e. $R=1$). And indeed, most (if not all) are likely to be innocent.

Scenario 2 (aka Advanced AI-empowered policing): $\forall d \in \{\textit{majority}, \textit{minority}\}, \forall u \in \mathbb{R}^l, \xi_{d,u} \simeq 0$, i.e., most reported individuals are criminals. This naturally arises in settings such as (not so futuristic) AI-empowered policing for various forms. Ideally, a good AI solution would highlight those incidents that are likely to have higher chance of crime. That is, if a good AI-system reports an incident (i.e. $R=1$), then the chance of crime is likely to be higher. However, such usage of AI solutions in policing should always be cautioned and heavily monitored.

Scenario 3 (aka Police-civilian interactions): $\forall d \in \{\textit{majority}, \textit{minority}\}, \forall u \in \mathbb{R}^l, 0 < \xi_{d,u} < 1$, i.e., a reasonable fraction of reported individuals are criminals as well as innocents. This naturally arises in the settings such as police-civilian interaction, including traffic stops and stop-frisk. Specifically, a good fraction of incidences reported through 911 calls involve innocents.

In what follows, we discuss how in each of these scenarios, under natural conditions how the primary source of racial disparity can be identified if $\Delta(u) \neq 0$ for some $u \in \mathbb{R}^l$. Towards that, we derive a characterization that would be very useful.

3.1 Observational Racial Disparity: A Useful Characterization

We present an alternate characterization of $\Delta(u)$ for $u \in \mathbb{R}^l$ involving $\xi_{d,u}$ with $d \in \{\text{majority}, \text{minority}\}$. Towards that, we define the following: for $u \in \mathbb{R}^l$ and $d \in \{\text{majority}, \text{minority}\}$,

$$\begin{aligned}Y_{\text{cri},d,u} &= \mathbb{P}[Y=1, M=1 \mid R=1, C=1, D=d, U=u], \\ Y_{\text{inn},d,u} &= \mathbb{P}[Y=1, M=1 \mid R=1, C=0, D=d, U=u].\end{aligned} \tag{12}$$

The above represent the probability of criminals and innocents, respectively, being subject to law enforcement actions.

Theorem 2. *For any $u \in \mathbb{R}^l$, the observational racial disparity in law enforcement can be represented as*

$$\Delta(u) = \Big(Y_{inn,maj,u} \cdot \xi_{maj,u} + Y_{cri,maj,u} \cdot (1 - \xi_{maj,u})\Big) - \Big(Y_{inn,min,u} \cdot \xi_{min,u} + Y_{cri,min,u} \cdot (1 - \xi_{min,u})\Big). \tag{13}$$

The proof of Theorem 2 can be found in Appendix Section D.

3.2 Scenario 1 (Aka Airport Security Checks)

As discussed earlier, in the setting such as airport security checks, where individuals self-report themselves and are later checked by TSA officers, it is reasonable to assume that most people passing through these security checks are innocent, not criminals. In our notation, this suggests that no matter what the context u and race d is, $\xi_{d,u} \simeq 1$. Therefore, by characterization of (13), we have $\Delta(u) \simeq Y_{\text{inn,maj},u} - Y_{\text{inn,min},u}$. Formally, we state the following.

Proposition 2. *For each $d \in \{majority, minority\}$ and $u \in \mathbb{R}^l$, let $\xi_{d,u} \to 1$. Then,*

$$\Delta(u) < 0 \Leftrightarrow Y_{inn,min,u} > Y_{inn,maju}. \tag{14}$$

Proof can be found in Appendix Section E.

Key takeaway: in scenario 1, if there is observational disparity against the minority (resp. majority) in law enforcement, it is primarily due to bias in law enforcement actions against the minority (resp. majority) innocents. That is, $\Delta(u) < 0 \Leftrightarrow Y_{\text{inn,min},u} > Y_{\text{inn,maj}u}$.

3.3 Scenario 2 (Aka Advanced AI-Empowered Policing)

Consider a (not so futuristic) scenario where the reporting system is enhanced with physical sensors and detectors that use accurate artificial intelligence (AI) [3]. For example, an advanced gunshot detection system can use AI to recognize a blast received by the acoustic sensors as a gunshot and then pinpoint its exact location to alert the police about this potential gun-related violation or crime. Similarly, a face recognition system that implements Computer Vision algorithms to recognize wanted criminals from the street surveillance footage can report them to the police. With the help of AI, these systems often accurately identify criminals, especially those who have committed felonies. In our notation, this suggests that no matter what the context u and race d is, $\xi_{d,u} \simeq 0$. Therefore, by characterization of (13), we should have $\Delta(u) \simeq Y_{\text{cri,maj},u} - Y_{\text{cri,min},u}$. Formally, we state the following.

[3] We do not intend to advocate for the implementation of AI technologies in law enforcement. Also, we refer to "AI" in a broader, more generalizable context. For example, it's akin to using automated systems like motion-sensor lights that automatically turn off after five minutes of inactivity in unoccupied conference rooms.

Proposition 3. *For each $d \in \{majority, minority\}$ and $u \in \mathbb{R}^l$, let $\xi_{d,u} \to 0$. Then,*

$$\Delta(u) < 0 \Leftrightarrow Y_{cri,min,u} > Y_{cri,maju}. \tag{15}$$

Proof can be found in Appendix Section F.

Key takeaway: in scenario 2, if there is observational disparity against the minority (resp. majority) in law enforcement, it is primarily due to bias in law enforcement actions against minority (resp. majority) criminals.

3.4 Scenario 3 (Aka Everyday Police-Civilian Interactions)

Now consider a scenario, such as police-civilian interactions, where the reporting is initiated through 911 calls. In this case, a substantial fraction of reporting corresponds to some form of criminal activity, but there is also a non-trivial amount of human error. That is, $\xi_{d,u}$ is neither close to 0, nor close to 1. Understanding the source of observational disparity can become rather challenging in this setting. There are three factors to consider: (a) the accuracy of the public 911 reporting, (b) policing actions on criminals, and (c) policing actions on innocents. Imposing restrictions on any two of these three factors can help us understand the other factor's impact in inducing observational disparity. We do that next.

Case 1. Consider the setting where public reporting is unbiased and policing actions against criminals are unbiased. That is, the 911-reported civilians of different races are equally likely to be innocents ($\xi_{\text{maj},u} \simeq \xi_{\text{min},u}$), and criminals of different races are equally likely to be subject to policing actions ($Y_{\text{cri,maj},u} \simeq Y_{\text{cri,min},u}$). Therefore, by characterization of (13), we have $\Delta(u) \simeq \xi_{\text{maj},u}(Y_{\text{inn,maj},u} - Y_{\text{inn,min},u})$. Formally, we state the following.

Proposition 4. *For each $d \in \{majority, minority\}$ and $u \in \mathbb{R}^l$, let $\xi_{maj,u} - \xi_{min,u} \to 0$ and $Y_{cri,maj,u} - Y_{cri,min,u} \to 0$. Then,*

$$\Delta(u) \to \xi_{maj,u}(Y_{inn,maj,u} - Y_{inn,min,u}). \tag{16}$$

Proof can be found in Appendix Section G.

Key takeaway: in scenario 3 case 1, there is observational disparity against the minority in law enforcement if and only if the minority innocents are more often subject to policing actions.

Case 2. Consider the setting where public reporting is unbiased and policing action for innocents is unbiased. That is, the 911-reported civilians of different races are roughly equally likely to be innocents ($\xi_{\text{maj},u} \simeq \xi_{\text{min},u}$), and innocents of different races are equally likely to be subjected to policing actions ($Y_{\text{inn,maj},u} \simeq Y_{\text{inn,min},u}$). Therefore, by characterization of (13), we have $\Delta(u) \simeq (1 - \xi_{\text{maj},u})(Y_{\text{cri,maj},u} - Y_{\text{cri,min},u})$. Formally, we state the following.

Proposition 5. *For each $d \in \{majority, minority\}$ and $u \in \mathbb{R}^l$, let $\xi_{maj,u} - \xi_{min,u} \to 0$ and $Y_{inn,maj,u} - Y_{inn,min,u} \to 0$. Then,*

$$\Delta(u) \to (1 - \xi_{maj,u})(Y_{cri,maj,u} - Y_{cri,min,u}). \tag{17}$$

Proof can be found in Appendix Section H.

Key takeaway: in scenario 3 case 2, there is observational disparity against the minority in law enforcement if and only if the minority criminals are more often subject to policing actions.

Case 3. Consider the setting where policing action for innocents and criminals is unbiased. That is, innocents and criminals of different races are equally likely to be subjected to policing actions ($Y_{\text{inn,maj},u} \simeq Y_{\text{inn,min},u}$ and $Y_{\text{cri,maj},u} \simeq Y_{\text{cri,min},u}$). Therefore, by characterization of (13), we have $\Delta(u) \simeq (Y_{\text{cri,maj},u} - Y_{\text{inn,maj},u})(\xi_{\text{min},u} - \xi_{\text{maj},u})$. Formally, we state the following.

Proposition 6. *For* $d \in \{majority, minority\}$ $u \in \mathbb{R}^l$*, let* $Y_{inn,maj,u} - Y_{inn,min,u} \to 0$*,* $Y_{cri,maj,u} - Y_{cri,min,u} \to 0$*. Then,*

$$\Delta(u) \to (Y_{cri,maj,u} - Y_{inn,maj,u})(\xi_{min,u} - \xi_{maj,u}). \tag{18}$$

Proof can be found in Appendix Section I.

Key takeaway: in scenario 3 case 3, there is observational disparity against the minority in law enforcement if and only if the 911-reported majorities are more likely to be innocents. That is, observational bias against the minority is due to actual reporting bias against the majority race!

4 Empirics

In this section, we utilize publicly available datasets on calls for service (i.e. 911 calls) and police stops from NYC and New Orleans to evaluate the implications of our framework. At its core, our goal is to understand whether there is evidence of bias either in law enforcement or reporting or both given the observed data. To that end, Section 2 provides a data-driven way to evaluate racial disparity. In particular, Proposition 1 as well Propositions 4-6 provide means to interpret the observational racial disparity.

Test Statistics and Interpretation. Given context $u \in \mathbb{R}^l$, the test statistics is $\Delta(u)$, the observational disparity in law enforcement. Using available data, for each $u \in \mathbb{R}^l$, we test the hypotheses

$$H_0(u) : \Delta(u) = 0 \text{ vs. } H_1(u) : \Delta(u) \neq 0 \tag{19}$$

Specifically, we compute the 95% confidence interval for the value of $\Delta(u)$, and if the interval does not contain 0, then we reject $H_0(u)$. In that case, per Proposition 1, we conclude presence of racial disparity (of some form).

If we conclude presence of racial disparity, to pin-point the cause further, we utilize Propositions 4-6.

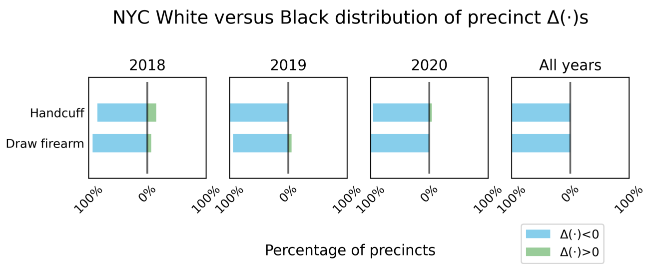
- *Case 1.* If public reporting and policing actions with respect to criminals are unbiased, then
  - $\Delta(u) < 0$ is primarily due to the bias against the minority innocents in policing actions.
  - $\Delta(u) > 0$ is primarily due to the bias against the majority innocents in policing actions.
- *Case 2.* If public reporting and policing action with respect to innocents are unbiased, then
  - $\Delta(u) < 0$ is primarily due to the bias against minority criminals in policing actions.
  - $\Delta(u) > 0$ is primarily due to the bias against majority criminals in policing actions.
- *Case 3.* If policing action with respect to innocents and criminals are unbiased, then
  - $\Delta(u) < 0$ is primarily due to the bias against the majority in public reporting.
  - $\Delta(u) > 0$ is primarily due to the bias against the minority in public reporting.

Specifically, given the context, we shall utilize one of the three cases above to conclude appropriate likely primary cause for racial disparity.

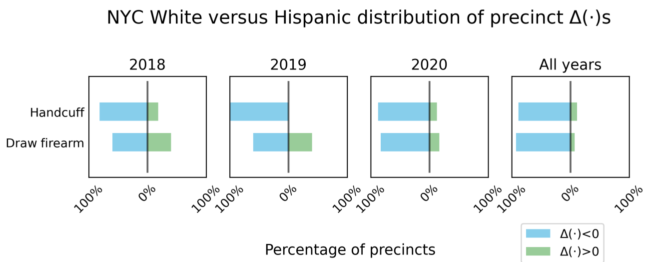
Details of Empirical Study. Various details related to our empirical study are provided in a thorough manner in Appendix Section J.3 due to space constraints. However, we highlight the choice of a proxy for the context U here. Whether all confounders have been accounted for and how to best account for them, especially when they are not fully observable, are the intrinsic challenges for any work in the field of causal inference with observational data. To define the *context* (i.e. U) in a measurable way given the data availability constraints, we utilize the precinct as a proxy following the standard of the field (Gelman, Fagan, and Kiss 2007). This is because a precinct often (1) consists of a comparatively homogeneous neighborhood in terms of the socioeconomic status and (2) is served by the same batch of police officers for relatively consistent law enforcement. To validate our results using the precinct as a proxy, we also conducted the same analysis using block groups in place of precincts, which include 600 to 3000 residents and represent the smallest geographical unit for which the bureau publishes a full range of demographic statistics, thus making them much more granular than the precincts. With this proxy choice of U, the outcome remains qualitatively unchanged, which gives us enough confidence in our proxy choice in this study. The details about data and associated necessary data processing related to the 911 call, police stop and associated use of force are explained in Appendix Section J.3. Next we describe our findings for NYC and New Orleans, and our likely conclusions.

4.1 New York City (NYC)

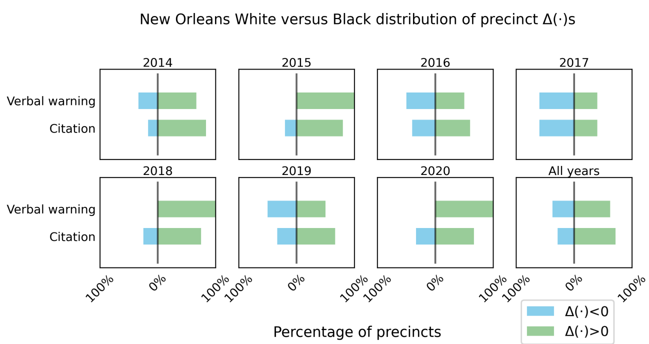
Data. NYC police commands exercise their authority over 77 police precincts. We utilize the 911 call data and pedestrian stop data over the years 2018, 2019, and 2020. We consider *White* as the majority race while *Black* or *Hispanic* as the minority race in the data analysis. After appropriate data processing, it contains $19,509,773$ total 911 calls-for-service, $34,009$ stops among which $21,130$ are induced from 911 calls. In this sense, 0.11% of 911 calls result in stops and 62.13% are induced by 911 calls. This suggests that only a small portion of 911 calls end up in stops but the majority of stops are from 911 calls. It stresses the importance of 911 calls in police-civilian encounters. The detailed year-to-year data statistics are summarized in Table 2 in Appendix Section J.



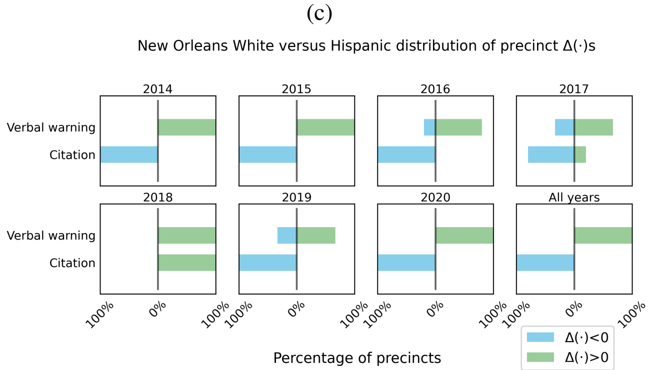
(a)



(b)



(c)



(d)

Figure 3: Distribution of statistically significant precinct-level $\Delta(\cdot)$s for different policing actions. $\Delta(\cdot) < 0$ indicates bias against minority (Black/Hispanic); $\Delta(\cdot) > 0$ indicates bias against majority (White). (a) NYC White versus Black; (b) NYC White versus Hispanic; (c) New Orleans White versus Black; (d) New Orleans White versus Hispanic. In summary, there is predominantly observation bias against minority (Black / Hispanic) compared to majority (White) in NYC while in New Orleans there is predominantly observation bias against majority (White) compared to minority (Black / Hispanic). We conclude that the primary source of bias in NYC is policing action against minority, while in New Orleans it is in the 911 calling (reporting) against minority!

In keeping with (Fryer 2019), use of force in a police-civilian interaction is classified at different levels with an additional level 0 corresponding to *no use of force*. A police-civilian interaction may involve several different levels of use of force and we associate the incident with the highest level of force used.

Key Takeaway. We obtain the results of $\Delta(\cdot)$s for NYC across all precincts as summarized in Figure 3 (a) and (b). The blue bars and green bars represent the proportions of precincts with statistically significant negative $\Delta(\cdot)$s and positive $\Delta(\cdot)$s, i.e. reject the null hypothesis (19).

We see that across different levels of use of force, it is comparatively consistent that most statistically significant $\Delta(\cdot)$'s are negative, exhibiting observational bias in law enforcement against the minority. In the case of NYC, it seems that Case 1 or Case 2 discussed above is likely. That is, the reason for observed racial disparity against minority is primarily due to *bias in policing action against innocents or criminals of the minority race*.

4.2 New Orleans

Data. New Orleans is divided into eight police districts. After data pre-processing, from 2014 to 2020, there are $2,160,679$ total 911 calls-for-service and $349,942$ stops. Of these stops, $134,746$ are induced from 911 calls. That is, 6.24% of 911 calls result in stops and 38.51% of stops are induced from by 911 calls. The detailed year-to-year statistics of data is summarized in Table 3 in Appendix Section J.

Key Takeaway. The summary findings for New Orleans are shown in Figure 3 (c) and (d). The blue bars and green bars represent the proportions of precincts with statistically significant negative $\Delta(\cdot)$s and positive $\Delta(\cdot)$s. As can be seen, unlike NYC, for New Orleans, we consistently find most statistically significant $\Delta(\cdot)$'s are positive, exhibiting observational bias in law enforcement *against the majority*. The block group level results remain qualitatively unchanged compared to precinct level results (see Appendix Section J.2).

Now if we consider Cases 1 or 2, then it would conclude that the police action is biased against majority which is not a likely scenario. Given that, the Case 3 is more likely. In which case, we would conclude that the primary source for *observational bias against majority is the bias against minorities (through over-reporting) in 911 calls-for-service.*

5 Conclusion and Future Work

In this work, we have developed a causal framework to evaluate racial bias or disparity in a multi-stage law enforcement system. We furbished it with data-driven test to detect the presence of bias. We identified three canonical scenarios where it is feasible to identify the primary source of bias when it is present. We applied our framework to the setting of police-civilian interaction using police stop data and 911 call data from NYC and New Orleans. We found that there is observational racial disparity against the minority in NYC while against the majority in New Orleans. Our framework leads to the conclusion that the likely source of bias in NYC is the biased policing actions against the minority while the likely source of bias in New Orleans is the biased 911 call reporting against the minority.

We highlight that discerning the potential source of racial bias in law enforcement systems is not to assign blame but to inform effective policy reforms. It is important to note that our conclusions are derived from empirical analyses using data collected over a certain period of time. Therefore, we advocate that any proposed policy changes should be grounded with the latest data to ensure that reforms are relevant, timely, and based on the most current understanding of the issues at hand.

Our work advances state-of-art in multiple ways. First, it provides a comprehensive framework by including notion of criminality leading to a proper notion racial (dis)parity, and thus overcoming the seeming debate in the prior works of Knox, Lowe, and Mummolo (2020) and Gaebler et al. (2020) over the right definition of causal estimand to determine the racial (dis)parity. Second, it provides a multi-stage framework to capture the entire causal chain of interactions. Third, our work provides data-driven test to detect the racial disparity. Fourth, through theoretical analysis, we identify canonical scenarios where it is feasible to identify the primary source of bias in the presence of it. Finally, the empirical analysis using police stop data along with 911 calls data provides new insights into the study of racial bias that has received extensive attention in the recent literature.

Our framework has applicability beyond the police-civilian interaction. For example, its applicability in the context of AI-empowered policing with sharp theoretical understanding would be extremely useful for evaluating the biases within AI-empowered systems as they get deployed more and more in the upcoming future. For instance, a similar causal framework can be utilized to evaluate bias within AI-empowered recruitment process. In particular, a recruitment process is a multi-stage system where a candidate of a particular race must first have access to a job opening information to apply, be interviewed, and then either be hired or not. Minority candidates with qualifications may already be at a disadvantage compared to majority candidates even before the actual interviews if they have less access to job opening information. And if an AI-empowered system is utilized to decide on whether to interview a candidate or not based on their resumes and job application, then it starts looking very similar to the scenario of AI-empowered policing we discussed earlier.

References

Akpinar, N.; De-Arteaga, M.; and Chouldechova, A. 2021. The effect of differential victim crime reporting on predictive policing systems. *CoRR*, abs/2102.00128.

Alesina, A.; and La Ferrara, E. 2014. A Test of Racial Bias in Capital Sentencing. *American Economic Review*, 104(11): 3397–3433.

Antonovics, K.; and Knight, B. G. 2009. A New Look at Racial Profiling: Evidence from the Boston Police Department. *The Review of Economics and Statistics*, 91(1): 163–177. Publisher: The MIT Press.

Blair, G.; Weinstein, J. M.; Christia, F.; Arias, E.; Badran, E.; Blair, R. A.; Cheema, A.; Farooqui, A.; Fetzer, T.; Grossman, G.; Haim, D.; Hameed, Z.; Hanson, R.; Hasanain, A.; Kronick, D.; Morse, B. S.; Muggah, R.; Nadeem, F.; Tsai, L. L.; Nanes, M.; Slough, T.; Ravanilla, N.; Shapiro, J. N.; Silva, B.; Souza, P. C. L.; and Wilke, A. M. 2021. Community policing does not build citizen trust in police or reduce crime in the Global South. *Science*, 374(6571).

Blumstein, A.; and Cohen, J. 1979. Estimation of Individual Crime Rates from Arrest Records. *The Journal of Criminal Law and Criminology (1973-)*, 70(4): 561–585.

Chouldechova, A. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.

Elwert, F.; and Winship, C. 2014. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40: 31–53.

Fryer, R. G. 2019. An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy*, 127(3): 1210–1261.

Gaebler, J.; Cai, W.; Basse, G.; Shroff, R.; Goel, S.; and Hill, J. 2020. A Causal Framework for Observational Studies of Discrimination. *arXiv:2006.12460 [stat]*. ArXiv: 2006.12460.

Gelman, A. 2020a. Statistical controversy on estimating racial bias in the criminal justice system.

Gelman, A. 2020b. Update on social science debate about measurement of discrimination.

Gelman, A.; Fagan, J.; and Kiss, A. 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 association*, 102(479): 813–823.

Goff, P. A.; and Kahn, K. B. 2012. Racial Bias in Policing: Why We Know Less Than We Should: Policing Racism. *Social Issues and Policy Review*, 6(1): 177–210.

Goncalves, F.; and Mello, S. 2021. A Few Bad Apples? Racial Bias in Policing. *American Economic Review*, 111(5): 1406–1441.

Greiner, D. J.; and Rubin, D. B. 2011. CAUSAL EFFECTS OF PERCEIVED IMMUTABLE CHARACTERISTICS. *The Review of Economics and Statistics*, 93(3): 775–785.

Grogger, J.; and Ridgeway, G. 2006. Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness. *Journal of the American Statistical Association*, 101(475): 878–887.

Hardt, M.; Price, E.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hellman, D. 2020. Measuring Algorithmic Fairness. *Virginia Law Review*, 106(4).

Hoekstra, M.; and Sloan, C. 2022. Does Race Matter for Police Use of Force? Evidence from 911 Calls. *American Economic Review*, 112(3): 827–60.

Horel, T.; Masoero, L.; Agrawal, R.; Roithmayr, D.; and Campbell, T. 2021. The CPD Data Set: Personnel, Use of Force, and Complaints in the Chicago Police Department. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Huang, X.; and Ning, J. 2012. Analysis of multi-stage treatments for recurrent diseases. *Statistics in Medicine*, 31(24): 2805–2821. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5456.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding Discrimination through Causal Reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 656–666. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR*, abs/1609.05807.

Knowles, J.; Persico, N.; and Todd, P. 2001. Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy*, 109(1): 203–229. Publisher: The University of Chicago Press.

Knox, D.; Lowe, W.; and Mummolo, J. 2020. Administrative Records Mask Racially Biased Policing. *American Political Science Review*, 114(3): 619–637.

Kramer, R.; and Remster, B. 2018. Stop, Frisk, and Assault? Racial Disparities in Police Use of Force During Investigatory Stops. *Law & Society Review*, 52(4): 960–993. Publisher: [Wiley, Law and Society Association].

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4069–4079. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Mitchell, O.; and Caudy, M. S. 2015. Examining Racial Disparities in Drug Arrests. *Justice Quarterly*, 32(2): 288–313. Publisher: Routledge _eprint: https://doi.org/10.1080/07418825.2012.761721.

Neil, R.; and Winship, C. 2019. Methodological Challenges and Opportunities in Testing for Racial Discrimination in Policing. *Annual Review of Criminology*, 2(1): 73–98.

Nix, J.; Campbell, B. A.; Byers, E. H.; and Alpert, G. P. 2017. A Bird's Eye View of Civilians Killed by Police in 2015 Civilians Killed by Police: Research Article. *Criminology and Public Policy*, 16(1): 309–340.

Pearl, J. 2014. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4): 459–481.

Pierson, E.; Simoiu, C.; Overgoor, J.; Corbett-Davies, S.; Jenson, D.; Shoemaker, A.; Ramachandran, V.; Barghouty, P.; Phillips, C.; Shroff, R.; and Goel, S. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(7): 736–745. Number: 7 Publisher: Nature Publishing Group.

Ridgeway, G. 2006. Assessing the Effect of Race Bias in Post-traffic Stop Outcomes Using Propensity Scores. Technical report, RAND Corporation.

Ridgeway, G.; et al. 2007. Analysis of racial disparities in the New York Police Department's stop, question, and frisk practices.

Sen, M.; and Wasow, O. 2016. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. *Annual Review of Political Science*, 19(1): 499–522. _eprint: https://doi.org/10.1146/annurev-polisci-032015-010015.

Simoiu, C.; Corbett-Davies, S.; and Goel, S. 2017. The Problem of Infra-marginality in Outcome Tests for Discrimination. arXiv:1607.05376.

Skogan, W.; and Frydl, K. 2004. Fairness and Effectiveness in Policing: The Evidence. Technical report, National Research Council, Washington, DC.

Slough, T. 2022. Phantom Counterfactuals.

Verma, S.; and Rubin, J. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 9781450357463.

Voigt, R.; Camp, N. P.; Prabhakaran, V.; Hamilton, W. L.; Hetey, R. C.; Griffiths, C. M.; Jurgens, D.; Jurafsky, D.; and Eberhardt, J. L. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25): 6521–6526.

Weisburst, E. K. 2022. "Whose help is on the way?": The importance of individual police officers in law enforcement outcomes. *Journal of Human Resources*, 0720–11019R2.

White, A. 2019. Misdemeanor Disenfranchisement? The Demobilizing Effects of Brief Jail Spells on Potential Voters. *American Political Science Review*, 113(2): 311–324.