Copyright 2019 by Issa Kohler-Hausmann

Printed in U.S.A.
Vol. 113, No. 5

EDDIE MURPHY AND THE DANGERS OF COUNTERFACTUAL CAUSAL THINKING ABOUT DETECTING RACIAL DISCRIMINATION

Issa Kohler-Hausmann

ABSTRACT—The model of discrimination animating some of the most common approaches to detecting discrimination in both law and social science—the counterfactual causal model—is wrong. In that model, racial discrimination is detected by measuring the "treatment effect of race," where the treatment is conceptualized as manipulating the raced status of otherwise identical units (e.g., a person, a neighborhood, a school). Most objections to talking about race as a cause in the counterfactual model have been raised in terms of manipulability. If we cannot manipulate a person's race at the moment of a police stop, traffic encounter, or prosecutorial charging decision, then it is impossible to detect if the person's race was the sole cause of an unfavorable outcome. But this debate has proceeded on the wrong terms. The counterfactual causal model of discrimination is not wrong because we can't work around the practical limits of manipulation, as evidenced by both Eddie Murphy's comic genius in the *Saturday Night Live* skit "White Like Me" and the entire genre of audit and correspondence studies. It is wrong because to fit the rigor of the counterfactual model of a clearly defined treatment on otherwise identical units, we must reduce race to only the signs of the category, meaning we must think race *is* skin color, or phenotype, or other ways we identify group status. And that is a concept mistake if one subscribes to a constructivist, as opposed to a biological or genetic, conception of race. The counterfactual causal model of discrimination is based on a flawed theory of what the category of race references, how it produces effects in the world, and what is meant when we say it is wrong to make decisions of import *because of* race. I argue that DISCRIMINATION is a thick ethical concept that at once describes and evaluates the actions to which it is applied, and therefore, we cannot detect actions as discriminatory by identifying a relation of counterfactual causality; we can do so only by reasoning about the action's distinctive wrongfulness by referencing what constitutes the very categories that are the objects of concern. An adequate theory of discrimination must rest upon (1) an account of the system of social meanings or practices that constitute the

categories at issue and (2) a moral theory of what is fair and just in various state and private arenas *given* what the categories are.

AUTHOR—Associate Professor of Law and Sociology, Yale University. I am deeply appreciative of Gabi Abend, Khiara Bridges, Philip Goodman, Scott Shapiro, and especially Tracey Meares for reading early and messy drafts of these ideas and providing incredibly valuable feedback and encouragement, as well as to Andrew Koppelman for supporting my nascent academic pursuits on this topic during my first year of law school. Diane de Gramont and Kevin Tobia provided excellent research assistance and substantive input. I also thank Julius Adebayo, Bruce Ackerman, Jack Balkin, Max Besbris, Deirdre Bloom, Owen Fiss, Jennifer Herwig, Daniel Markowitz, John Levi Martin, Osagie K. Obasogie, Robert Post, Mike Seidman, Andrew Selbst, Reva Siegel, Andrew Smart, Michael Carl Tschantz, Petri Ylikoski, and Tukufu Zuberi for generous and engaged discussion on the project. Special gratitude is due to Chris Winship for reading drafts and indulging in extended email exchange about these topics, and to Jim Greiner, whose careful and substantial engagement with me about these ideas despite our disagreements is the mark of a generous person and a dedicated intellectual. I am grateful to the student editors at the *Northwestern University Law Review* for taking a chance on an unorthodox piece of scholarship for their inaugural issue dedicated to empirical legal studies and for their careful edits and thoughtful feedback. Immense gratitude is due to Gideon Yaffe, without whose guidance, feedback, and patient emotional support this project would not have moved forward. I dedicate this Article to the memory of two inspirational people who passed in 2018: Devah Pager, whose passion for racial justice inspired a brilliant and impactful academic career, and Mujahid Farid, whose tireless advocacy for humane parole reform changed the system and all of us who had the opportunity to work with him.

INTRODUCTION: DEFINING AND DETECTING DISCRIMINATION

Judge Schroeder did not believe Dr. Lamberth could pick out Hispanic drivers by looking at them.[1]

To be more specific, Judge Schroeder did not find the method Dr. Lamberth employed to create a "benchmark" of the objective rate at which Hispanic drivers violated traffic laws in North Carolina's Alamance County to be scientific.

To be even more specific, Judge Schroeder did not believe that Dr. Lamberth had deployed an objective, replicable, or verifiable method to detect the rate at which Hispanic drivers violated North Carolina traffic laws by hiring two auditors, Mr. Rivera and Mr. Valdez, to sit in parked cars on select roadways, observe passing cars, count which were violating North Carolina traffic laws, and look at drivers to see "who 'appeared to be' or 'looked' Hispanic."[2]

Judge Schroeder noted that other cases and other peer-reviewed studies relied upon a similar observational methodology to construct a benchmark of the rate at which a designated demographic group violated certain laws. However, he pointed out that those studies "utilized more reliable methods of observation," and were comparing "African-American drivers, not Hispanic drivers, to non-African-American drivers," which, his reasoning implied, presented obvious and unproblematic indicia of racial status.[3] Judge Schroeder concluded that because "no control, standard, or description was used to identify Hispanics[,] . . . Dr. Lamberth offered no information on what, if any, standard [the auditors] used," and "Dr. Lamberth's study thus relies entirely on the subjective views of Rivera and Valdez and their

[1] United States v. Johnson, 122 F. Supp. 3d 272, 331 (M.D.N.C. 2015). Debates about the visual obviousness of racial and ethnic categories is a longstanding tradition in American law. See, for example, the fascinating discussion of the racial status of Hindus in *United States v. Bhagat Singh Thind*, 261 U.S. 204 (1923), in Sherally Munshi, *"You Will See My Family Became So American": Toward a Minor Comparativism*, 63 AM. J. COMP. L. 655, 656 (2015), or the discussion of the adjudication of the obviousness of Alice Jones's blackness in Angela Onwuachi-Willig, *A Beautiful Lie: Exploring* Rhinelander v. Rhinelander *as a Formative Lesson on Race, Identity, Marriage, and Family*, 95 CALIF. L. REV. 2393, 2399 (2007).

[2] *Johnson*, 122 F. Supp. 3d at 305.

[3] *Id.* at 332.

personal, totally subjective say-so of who should be considered 'Hispanic.'"[4] Judge Schroeder, therefore, excluded Dr. Lamberth's expert report and testimony as failing to meet the *Daubert* standards for admissibility, including testability, known error rates, peer review, and general acceptance in the scientific community.[5]

Based on his conclusion that this and another study could not prove discrimination, Judge Schroeder ruled in 2015 that the United States Department of Justice failed in its more-than-three-year effort to show that the Alamance County Sheriff engaged in "a pattern or practice of discriminatory law enforcement activities directed against Latinos in Alamance County" in violation of the Fourth and Fourteenth Amendments.[6]

How do we know when a particular act, practice, or policy is an instance of DISCRIMINATION?[7] What precisely do we mean when we identify discrimination as an act, practice, or policy taken "because of" race or ethnicity? This Article will probe these questions in one arena of social life in which this author just happens to have experience and interest—police and prosecutorial racial discrimination. The conceptual analysis offered here is applicable to other arenas of social life—from employment, to housing, to credit—and has modified implications for how to conceptualize other categories of discrimination—from sex to sexual orientation. However, I will stick to the example of race or ethnicity in criminal justice simply to focus the discussion and because these debates have immediate political salience as recent killings of unarmed black persons have pushed the question of

---

[4] *Id.* at 331.

[5] This is the court's "gatekeeping" function under FED. R. EVID. 702, and *Daubert* requires "that any and all scientific testimony or evidence admitted is not only relevant, but reliable." Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 589 (1993).

[6] Complaint at 1, *Johnson*, 122 F. Supp. 3d 272 (No. 1:12-cv-01349-TDS-JLW); *see also Johnson*, 122 F. Supp. 3d at 380 (concluding that the United States Department of Justice did not meet its burden of demonstrating that the Alamance County Sherriff's Office engaged in discriminatory law enforcement practices).

In 2012, the Department of Justice had filed suit under 42 U.S.C. § 14141, alleging that the Alamance County Sheriff's Office (ACSO), headed by Terry S. Johnson,

> disproportionately subjects Hispanics to unreasonable searches, arrests them for minor infractions (in lieu of issuing warnings or citations), targets them at vehicle checkpoints located in predominantly Hispanic neighborhoods, uses ethically-offensive epithets to refer to Hispanics and otherwise tolerates activities of deputies that evidence anti-Hispanic bias, automatically and selectively refers Hispanic arrestees to U.S. Immigration and Customs Enforcement ("ICE") investigators for deportation, and otherwise engages in deficient policies, training, and oversight that facilitates discriminatory enforcement.

Johnson, 122 F. Supp. 3d at 282.

[7] I use small caps for concepts (the concept DISCRIMINATION in noun form or DISCRIMINATORY in adjectival form); scare quotes for terms of art, expressions, or to indicate so-called usage; and italics for emphasis.

discriminatory police violence into mainstream public debates. So how do we detect discrimination when questions of great importance hang in the balance?

A. Discrimination as Outcomes Caused by Race

This Article argues that animating the most common approaches to detecting discrimination in both law and social science is a model of discrimination that is, well, wrong. I term this model the "counterfactual causal model" of race discrimination. Discrimination, on this account, is detected by measuring the "treatment effect of race," where treatment is conceptualized as manipulating the raced status of otherwise identical units (e.g., a person, a neighborhood, a school).[8] Discrimination is present when an adverse outcome occurs in the world in which a unit is "treated" by being raced—for example, black—and not in the world in which the otherwise identical unit is "treated" by being, for example, raced white.[9] The counterfactual model has the allure of precision and the security of seemingly obvious divisions or natural facts.[10] Despite notable objections,

---

[8] I will use the awkward terminology of "unit" throughout the Article, unless I am giving an example of particular types of units, in order to indicate that the model (and my objections to it) can encompass individual and aggregate units of analysis. The treatment would differ based on the type of unit. For example, we might imagine the individual-level treatment to be discrete raced status (e.g., white vs. black), and the treatment for aggregate-level units might be a continuous measurement (e.g., population composition measure). Although there are important differences between individual-level and aggregate-level units in terms of expressing the counterfactual model and the types of objections one could raise to the model as a conceptualization of discrimination, this Article is mostly dedicated to laying out the broad strokes of my objections.

[9] Levi Martin and King-To Yeung launch their exploration of the use of the category of race in sociology over sixty years with the following parable:

> There is an old Zen koan in which the master Shuzan Osho held up his staff before his disciples and said, "You monks! If you call this a staff, you oppose its reality. If you do not call it a staff, you ignore the fact. Tell me, you monks, what will you call it?" The discomfort felt by the monks, who had to choose between denying their insight into the fundamental oneness of the universe and making the absurd counterfactual denial of self-evident fact, is also felt by many sociologists when it comes to the analysis of race.

John Levi Martin & King-To Yeung, *The Use of the Conceptual Category of Race in American Sociology, 1937–99*, 18 SOC. F. 521, 521 (2003). In a rare moment of commonality with monks, I am similarly tortured to find terminology that at once acknowledges the socially constructed nature of race but also recognizes that in its current constructed form, it presents as a solid, obvious, and commonsensical "category of practice." Rogers Brubaker & Frederick Cooper, *Beyond "Identity,"* 29 THEORY & SOC'Y 1, 4 (2000). Like the monks, I have not found an answer to how to acknowledge the taken-for-granted status of race and also its contingent, constructed, and contested meaning and content. So, I will waver back and forth, probably frustrating everyone with my terminology, sometimes using *race* unproblematically when I am talking about the counterfactual model, and sometimes using awkward terms like "raced status" to indicate a black box of ascriptive meaning.

[10] Osagie Obasogie calls the current hegemonic understanding *"race" ipsa loquitur*: "[the] notion that race is not only visually obvious but that its social salience, perceptibility, and visual significance

this remains the leading conception of discrimination in both law and social science.[11] But I contend that this model is wrong. It is wrong because it is based on a flawed theory of (1) what the concept RACE references and how it produces effects in the world, and (2) what we mean when we say it is bad to make important decisions "because of race."

Much of this Article is dedicated to making the negative case against the predominant counterfactual causal model of discrimination by arguing that it is incompatible with the constructivist theory of race, to which most (but not all) academics and judges say they subscribe.[12] In the process, I propose a radically different way of conceptualizing discrimination that uses two concepts largely unfamiliar in debates about discrimination. Although these concepts might at first blush seem challenging, I contend that they are essential to any plausible approach to discrimination.

Objections to talking about race as a cause in the counterfactual framework are usually raised in terms of *manipulability*. Candidates for causes in the counterfactual framework are limited to viable treatments to which a unit could be subjected at the time the outcome of interest might occur.[13] If one cannot manipulate a person's race at the moment of a police stop, traffic encounter, or prosecutorial charging decision, then it is impossible to detect if the person's race was the sole cause of an unfavorable outcome. But, as many have pointed out, we should not confuse empirical and theoretical objections. If one accepts that race or ethnicity is the type of thing that is properly conceptualized as an isolated manipulation on units that can otherwise remain the same units, then there are workarounds to the practical problems of actual manipulation. We can, for example, imagine presenting a police officer or prosecutorial decision-maker with candidates for an outcome bearing identical credentials and vary some indicia of the candidate's racial status in order to detect the treatment effect of race. This is the logic of what are called audit studies—a method illustrated brilliantly

---

stem from self-evident distinctions . . . ." OSAGIE K. OBASOGIE, BLINDED BY SIGHT: SEEING RACE THROUGH THE EYES OF THE BLIND 143–44 (2014).

[11] *See infra* Parts I–II.

[12] And if one is *not* a constructivist about race, then I don't see any way one can recognize discrimination as a wrong distinctive from, say, general distributive injustice or inefficiency. *See infra* text accompanying notes 16–20.

[13] Paul W. Holland, *Statistics and Causal Inference*, 81 J. AM. STAT. ASS'N 945, 959 (1986).

> An attribute cannot be a cause in an experiment, because the notion of *potential exposability* does not apply to it. The only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit. Statements of 'causation' that involve attributes as 'causes' are always statements of association between the values of an attribute and a response variable across the units in a population.

Id. at 955.

by Eddie Murphy in the 1984 *Saturday Night Live* skit "White Like Me," where he applies white face makeup to see how he is treated as a white man in New York City.[14] But my concerns with conceptualizing race as a treatment as a way to define discrimination are theoretical, not practical.

The problem with identifying discrimination with the treatment effect of race is that it misrepresents what race is and how it produces effects in the world, and concomitantly, what makes discrimination because of race a moral wrong. In the classic counterfactual causal inference framework, race can be a treatment on units only if manipulating it does not entail fundamental changes to other aspects of the unit. Thus, audit studies can be interpreted as detecting the treatment effect of race and race alone by changing some signifier of racial status on candidates only if the manipulation does not transform the unit into a different unit in ways necessarily relevant to interpreting an action as discriminatory.[15] But if the signifiers of racial categories fundamentally structure the interpretation and relevance of other characteristics or traits of the unit, then it is a mistake to talk about identical units that differ only by raced status. Yet, to talk about race as an isolated treatment on units reduces it to some set of signifiers that elicits outcomes in the world only as a psychological trigger or stimulus to disfavor. And to do so is a gross category mistake, at least if you subscribe to the *constructivist theory* of race.

A constructivist rejects the notion that racial categories in the United States are constituted by genetic or biological facts, and instead holds that what now seem like obvious, taken-for-granted categories of racial difference were constructed over hundreds of years of historical practices starting with chattel slavery and colonization. Categories such as "whiteness" and "blackness" were forged through social relations such as forced labor, colonization, immigration, and Jim Crow; they were contested and policed through many institutions including immigration laws, housing and education segregation, violent regulation of social and intimate relations, and hoarding of occupational and economic opportunity. These processes

[14] The skit's name was a play on the book by John Howard Griffin, a white journalist who took medication to darken his skin and recorded his experiences traveling in the 1950s Deep South. JOHN HOWARD GRIFFIN, BLACK LIKE ME (2010).

[15] "Signifier" and "signified" are terms from semiotics with contested meanings, but for purposes of this project a simple definition is sufficient. Without subscribing to all of Barthes's semiotics, I will use those classic terms—signifier and signified—to roughly mean the forms, signals, and material of expression (the signifier) and the content expressed, concept referenced, or meaning imparted by the former (the signified). Signifiers come to signify particular concepts or values because of social, cultural, or historical convention, not because of some natural relationship between the form of the signifier and its signified. ROLAND BARTHES, ELEMENTS OF SEMIOLOGY 39–42 (Annette Lavers & Colin Smith trans., 1977).

made certain aspects of physical appearance salient markers of social difference and reinforced their relevance over many types of interactions. Just because we can trace the historical processes by which these categories were constructed does not make them any less real with real effects; a wink or hoodie can have fatal consequences because of the status of race. We continue to live in a world in which most important institutions are racialized, meaning they play an active role in reproducing the significance of these very categories.

This—with extreme brevity—is what is commonly termed the social constructivist theory of race, which holds that the racial categories as we currently know them are not inevitable distinctions that naturally arise from objective biological differences, but instead are the product of contingent historical social processes.[16] Thus, for a constructivist, the term "race" cannot refer to an attribute, a genetically produced trait, or a signifier—level of melanin in skin, phenotype, distinctive names or speech—that people just have and thereby obviously belong to a designated racial group. The term references a complexly constituted social fact, whereby material and dignitary opportunities are organized such that certain physical and cultural signifiers become the salient markers of consequential cultural categories, and those categories are constituted by a constellation of social relations and meanings with a definite content and organization.[17] Race in America is, as Bonilla-Silva and Zuberi say, "a social system that uses skin color as the criterion for classification. . . . Racial stratification is real, but biology is not its root cause."[18] Although the constructivist view is now widely accepted in

[16] There are numerous rich (and in some respects competing) frameworks that theorize the processes by which racial status or groups are constructed. Constructivist positions are articulated in so many places it is hard to cite to them, but a few well-known examples in sociology include the following: MUSTAFA EMIRBAYER & MATTHEW DESMOND, THE RACIAL ORDER 49 (2015); MICHAEL OMI & HOWARD WINANT, RACIAL FORMATION IN THE UNITED STATES 103–36 (3d ed. 2015); Eduardo Bonilla-Silva, *The Essential Social Fact of Race*, 64 AM. SOC. REV. 899, 899 (1999); and Mara Loveman, *Is "Race" Essential?*, 64 AM. SOC. REV. 891, 891 (1999).

[17] The term "social fact" is drawn from Durkheim, for whom social facts "consist of manners of acting, thinking and feeling external to the individual, which are invested with a coercive power by virtue of which they exercise control over him." EMILE DURKHEIM, THE RULES OF SOCIOLOGICAL METHOD AND SELECTED TEXTS ON SOCIOLOGY AND ITS METHOD 21 (Steven Lukes ed., W.D. Halls trans., 2d ed. 2013). Eduardo Bonilla-Silva described race as "the essential social fact," and I follow him in saying that defining race as an individual-level trait makes no sense because it fails to recognize that those traits only have meaning in particular racialized systems of material and symbolic hierarchies. *See* Bonilla-Silva, *supra* note 16, at 899.

[18] Tukufu Zuberi & Eduardo Bonilla-Silva, *Toward a Definition of White Logic and White Methods*, *in* WHITE LOGIC, WHITE METHODS: RACISM AND METHODOLOGY 3, 10 (Tukufu Zuberi & Eduardo Bonilla-Silva eds., 2008).

academic circles, its implications are not appreciated by the predominant legal and social scientific approaches for detecting discrimination.[19]

One implication of the social constructivist theory is that race cannot be conceptualized as an isolated treatment in the counterfactual causal model, and accordingly, racial discrimination cannot be defined as the treatment effect of race. If we accept the constructivist theory of race, then we must reject attempts to detect racial discrimination that seek to isolate the causal effect of race alone because it rests on a sociologically incoherent conception of what race references and how it can cause a distinctive form of action called *discrimination.*

The first conceptual tool I use to proffer an alternative account of discrimination that is compatible with the constructivist theory of race is borrowed from moral philosopher Bernard Williams, who coined the phrase "thick ethical concept" for terms that simultaneously describe and evaluate the object to which they are applied.[20] Thin ethical concepts, such as BAD, OUGHT, or RIGHT, do not require "institutional and cultural presuppositions" in order to impart judgment.[21] To apply the terms properly, you do not need access to complex social facts, and to say that an action is BAD or RIGHT does not convey more information about the evaluated action beyond the moral valuation. Thick ethical concepts, on the other hand, such as RESPECTABILITY, CHIVALRY, or PIETY, do require complex social knowledge in order to be used and decoded. To invoke the term is to simultaneously represent the evaluated action as a particular kind of action—one that is only classifiable as such using a cultural repertoire and understandings about the functioning of a particular social world—and to impart judgment. That is, to morally evaluate an action with a thick ethical concept communicates information about *the way in which* the action is bad that relies on institutional and cultural facts.[22]

Discrimination is not a thin ethical concept that can be represented as "choosing + bad," "arresting + mean," or "prosecuting + irrational," because

---

[19] ANN MORNING, THE NATURE OF RACE: HOW SCIENTISTS THINK AND TEACH ABOUT HUMAN DIFFERENCE 10–23 (2011) (showing that the constructivist view is widely accepted among social scientists, but the biological conception of race is still commonly held by undergraduates and the public at large); *see also* Martin & Yeung, *supra* note 9, at 521–25 (showing that although the constructivist position is explicitly embraced by social scientists, many fail to operationalize it in any meaningful way in their research methodology).

[20] BERNARD WILLIAMS, ETHICS AND THE LIMITS OF PHILOSOPHY 155–56 (2006).

[21] Gabriel Abend, *Thick Concepts and the Moral Brain*, 52 EUR. J. SOC. 143, 147–48 (2011). One might need situated sociological knowledge to understand the subject of a sentence using thin moral concepts, such as "Polygamy is bad," but the evaluative component is not entailed in the way in which the object is described.

[22] *Id.* at 149–58.

we know that the Fourteenth Amendment, Title VII, Title IX, and countless other state and federal statutes are not about outlawing all bad, mean, or irrational forms of state or private action. Discrimination is a thick ethical concept that can only be comprehended with access to situated cultural knowledge about the relevant categories that make up a particular society's system of stratification and a normative critique of how those categories operate. In order for something to be discriminatory—instead of merely mean, random, or irrational—the act or policy must rely on meanings or facts that constitute the social category in ways that we morally disavow. Therefore, any discrimination-detecting exercise must proceed from some moral theory—often implicit—of what is fair or just in the face of how a particular society's stratification works through meanings and relations of its social types.

This brings me to the second conceptual tool I use to build an alternative to the counterfactual causal model of discrimination, that of constitutive explanation, which I argue accurately captures the type of claim made when something is labeled discriminatory. A constitutive claim accounts for the capacities of complex systems by reference to their constitutive elements: the parts and organization that make the system what it is.[23] To identify something as discrimination when it happened "because of" race or ethnicity is not to name a relation of counterfactual dependence defined as an outcome triggered by isolating and manipulating an individual trait. To identify something as discrimination when it happened because of race or ethnicity is to offer a constitutive claim that explains how an action or practice can be morally objectionable by virtue of the complex of social meanings and relations that constitute the social category. A constitutive claim unifies a set of disparate practices (choosing, excluding, promoting, demoting, arresting, jailing, beating, humiliating, killing) as morally problematic in the same way, namely by reference to how the action or policy engages the content of the socially constructed category.

Combining these two conceptual components yields a definition of discrimination as an action or practice that acts on or reproduces an aspect of the category in a way that is morally objectionable. It is a thick ethical concept that—to express the distinctive wrongfulness of the action vis-à-vis the category—must rest upon an account of the system of social meanings or practices that constitute the categories at issue.

The definition contains empirical and normative elements, both of which are black-boxed in this Article. The first black box must contain social ontology or practical anthropology, requiring us to identify and define the

[23] *See infra* Part I.

stratifying social types in a given society; the second black box must contain political and moral philosophy, requiring us to decide what is fair and just in various state and private arenas *given* what the categories are. The point of this Article is not to fill in those black boxes, but to explain why both elements are fundamental to any discrimination-detecting endeavor.

B. What to Expect and Why It Matters

Before proceeding, let me be clear that my aim here is not to criticize quantitative methods, audit studies, or legal strategy from any case or research program. My aim is to make a set of sociological and analytic points concerning the meaning of those studies.[24] My conceptual points lead to political–strategic ones. At the risk of being disowned by my materialist intellectual family, I will say that ideas matter. They especially matter in the legal field, where the way in which powerful legal actors conceptualize an issue has profound implications for what they do with their power.

Currently, many courts, experts, and commentators approach detecting discrimination as an exercise measuring the counterfactual causal effect of race-qua-treatment, looking for complex methods to strip away confounding variables to get at a solid state of race and race alone.[25] But what we are arguing about when we argue about whether or not statistical evidence provides proof of discrimination is precisely *what we mean* by the concept DISCRIMINATION. We are arguing about the social meaning of race and how it structures outcomes of interest. Similarly, what we are arguing about when we debate what variables ought to be controlled for or balanced on in a quantitative exercise to detect discrimination is what are the fair or just grounds for decision-making or resource allocation *in light of* what race is and how it operates. We ought to be clear about the nature of the debates we are having such that the driving issues are not obfuscated by claims of methodological rigor or objective scientism.

Because thick ethical concepts and constitutive explanations are significantly less familiar ways of approaching discrimination than the counterfactual causal model, Part I offers a primer so that the reader may sense that there *is* an alternative way of thinking about this issue before I proceed to my negative case. Part II turns, briefly, to showing just how prevalent the counterfactual causal conceptualization is in both law and

---

[24] Many people who use the terminology of causal effects of race or who interpret methods in a way that suggests race can be conceptualized as a treatment in the counterfactual model may embrace the constructivist account of race but use the language because it is expedient. I simply caution that "[w]hat we do may be more important than what we think and what we say." Martin & Yeung, *supra* note 9, at 539.

[25] *See infra* Part II.

social science. To fully comprehend my critique of the model, it is essential to present it with formal rigor, and Part III does so. The first two Sections of Part III explore critiques of talking about race as a treatment in the counterfactual causal inference literature and lay out the most widely accepted retort that has been offered to support the counterfactual causal model of discrimination. The third Section of Part III fleshes out my sociological objection to race-qua-treatment by exploring how attempting to isolate the treatment effect of race is at odds with a constructivist account of race, which rejects the view that racial categories "reflect natural, stable differences between human groups."[26]

My arguments are pitched at a fairly high level of abstraction in Parts I through III, and therefore, Part IV explores the cash value of these theoretical points by thinking through the nitty-gritty of design and interpretation of audit or correspondence studies, which are usually touted as the gold standard for causal inference. In this Part, I argue that audit studies certainly can produce evidence of discrimination, but they don't do so by virtue of isolating the treatment effect of race. Audit studies are often recognized as compelling evidence of discrimination because they instantiate widely shared moral convictions, namely that, at a minimum, persons in the designated social groups with the given set of credentials ought to elicit the same treatment. But, properly understood, audit studies produce evidence of discrimination in the same way that analysis of observational data or an individual encounter do: by relying on a constitutive claim about what race is to ground a moral claim about what is distinctively wrong about the act or practice.

The final Part of this Article does not offer a new set of clean, determinate doctrinal formulations of discrimination, nor list magic-bullet methods to detect it. As will become clear, one of my main claims throughout this Article is that it is impossible to do so without a prior moral–political philosophy of what justice requires in private and public domains *in light of* what racial and ethnic stratification is in America today, a substantial project for a different paper (or book).[27] However, I suggest that an upshot of the arguments I advance about the counterfactual causal model is that distinctions between disparate treatment and impact that have been advanced in terms of the former being caused exclusively by race and the latter being

---

[26] MORNING, *supra* note 19, at 18.

[27] For such eloquent theories and debates, see CHRISTOPHER J. LEBRON, THE COLOR OF OUR SHAME: RACE AND JUSTICE IN OUR TIME (2013); Tommie Shelby, *Race and Social Justice: Rawlsian Considerations*, 72 FORDHAM L. REV. 1697 (2004); Charles W. Mills, *Retrieving Rawls for Racial Justice?: A Critique of Tommie Shelby*, 1 CRITICAL PHIL. RACE 1 (2013); and CHARLES W. MILLS, BLACK RIGHTS/WHITE WRONGS: THE CRITIQUE OF RACIAL LIBERALISM (2017).

caused by something that is not-race, but correlated with race, are not conceptually tenable distinctions.[28] They are not tenable because these formulations only make sense if one defines race *as* the visual or social cues associated with the category: race is skin color, or is phenotype, or is one of the physical or social signifiers of the category. If one subscribes to the constructivist notion of race—in which signifiers come to be indicative of a status only through entrenched social practices—then it is nonsensical to talk about constitutive practices as somehow being race-neutral things. Of course, we can advance other distinctions between what should or should not be legally actionable discrimination, but we can't do so by relying on value-free notions of counterfactual causality.

Many scholars have compellingly argued that the law of equal protection ought to be interpreted as a principle of antisubordination, or that the purpose of antidiscrimination law more broadly ought to be understood as a project of remaking social meanings of historically marginalized groups.[29] I agree with their arguments. But I approach this debate from a new angle, backing out a theory of the category of race from the prevailing methods used to detect discrimination. I conclude that if one subscribes to the constructivist theory of the category of race, then it is incoherent to understand the legal proscription against discrimination as anything but a project to remake the very meanings of social categories (unless you don't want to distinguish discrimination from mere irrationality or idiosyncrasy, but then you have a different problem—which is to explain why the state's heavy coercive machinery should be concerned with some forms of classification and not others).[30]

Insofar as we (and I use the first-person plural pronoun to indicate I understand this Article as a part of conversation with activists on this front) are interested in transforming the social structures that systematically

---

[28] The same conceptual points could be used to query the distinction between what economists call taste-based and statistical discrimination. The classic distinction defines the former as a willingness to pay to associate with members of group A instead of B notwithstanding identical productive qualifications, and the latter as using membership in A or B as a proxy for other instrumentally rational capacities or qualifications. *See* KENNETH J. ARROW, SOME MODELS OF RACIAL DISCRIMINATION IN THE LABOR MARKET (1971); GARY S. BECKER, THE ECONOMICS OF DISCRIMINATION 13–18 (2d ed. 1971); Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659, 659–61 (1972). My point is that the former is an instance of DISCRIMINATION only when social conditions—some systematic social and economic differences that produce the salience of A-ness in opposition to B-ness—make the latter possible, systematic social and economic differences such that A-ness is available as a proxy for functionally relevant qualifications.

[29] *See infra* Conclusion.

[30] Some have argued that, as a descriptive matter, much of antidiscrimination law *has* devolved into merely a proscription against irrationality. *See, e.g.*, Robert Post, *Prejudicial Appearances: The Logic of American Antidiscrimination Law*, 88 CALIF. L. REV. 1, 14 (2000). My point is to critique such conflation from basic logical, sociological, and philosophical principles.

oppress and disadvantage minority communities, we must be incredibly attentive to the way discourses about race and ethnicity circulate and settle in the halls of power. In presenting quantitative evidence to courts, it is a mistake to talk as if we have gotten at the *true* effect of race by modeling it as a counterfactual treatment. In fact, to do so cements an already predominant and problematic understanding about race in public and legal discourse: one that is distressingly dehistoricized and desocialized.[31] Discourses have effects. Folks in positions of power—namely judges considering discrimination cases—make important decisions because they understand words and concepts in a particular way. And we should train our sights on trying to make them understand things in what, I contend, is the right way. Doing so will bring to the fore difficult political and moral judgments that are at the heart of debates about discrimination broadly.

At a minimum, I hope to start a dialogue with the community that provides evidentiary expert statistical services to litigants in discrimination cases about how that material is consumed and given meaning in courts. Social scientists can do more in the fight against discrimination than provide technical skills; they can also offer careful reflexive thinking that rejects folk, commonsense-sounding concepts of race. We must be vigilant to "uncover the hidden assumptions in [our] own scientific unconscious . . . ."[32]

I. PRIMER: THICK ETHICAL CONCEPTS AND CONSTITUTIVE EXPLANATIONS

Discrimination is, to borrow from Bernard Williams, a "thick ethical concept" that simultaneously describes and evaluates the object to which it is applied.[33] To know if something is DISCRIMINATION, one needs thick sociological and anthropological knowledge about what constitutes the

[31] Others have raised objections to talking about race as a variable that causes other variables, highlighting that the practice is complicit (or active) in reifying race as an essential trait. *See, e.g.*, TUKUFU ZUBERI, THICKER THAN BLOOD: HOW RACIAL STATISTICS LIE 29–31 (2001); Philip Goodman, *Race in California's Prison Fire Camps for Men: Prison Politics, Space, and the Racialization of Everyday Life*, 120 AM. J. SOC. 352 (2014); Tukufu Zuberi, *Deracializing Social Statistics: Problems in the Quantification of Race*, 568 ANNALS AM. ACAD. POL. & SOC. SCI. 172 (2000). I draw on these objections in Parts III and IV.

[32] EMIRBAYER & DESMOND, *supra* note 16, at 72.

[33] WILLIAMS, *supra* note 20, at 155–56 ("Many exotic examples of these can be drawn from other cultures, but there are enough left in our own: *coward*, *lie*, *brutality*, *gratitude*, and so forth. They are characteristically related to reasons for action. If a concept of this kind applies, this often provides someone with a reason for action, though that reason need not be a decisive one and may be outweighed by other reasons, as we saw with their role in practical reasoning in Chapter 1. Of course, exactly what reason for action is provided, and to whom, depends on the situation, in ways that may well be governed by this and by other ethical concepts, but some general connection with action is clear enough. We may say, summarily, that such concepts are 'action-guiding.'").

categories to which it is applied: how certain practices, acts, and understandings differentiate humans into distinct social kinds known as RACES or ETHNICITIES, and, conversely, how such constituted categories give meaning to acts and credentials. One also needs a normative theory about the fair and just basis for making decisions, which can be advanced only *in light of* the sociological or anthropological facts that make race or ethnicity a meaningful category whereby it is discriminatory—as opposed to just irrational or idiosyncratic—to act on it. Said another way, unless we have a moral critique of how such categories operate, we have no grounds to identify acts or practices as discriminatory.[34] How should we think about detecting such a thick ethical concept?

In my view, the only plausible account of what we are doing when we set out to detect discrimination is seeking "constitutive explanations," which in turn ground thick ethical claims. Constitutive explanations explain properties of a system "by appealing to their parts and their organization."[35] For example, one might say that a wine glass broke in response to the strike of a spoon because it was fragile.[36] The explanation references what *constitutes* fragility by "detail[ing] the relevant aspects of the object's molecular structure that make it fragile," and by so doing, we understand why, under "certain enabling and triggering conditions," the object breaks.[37] Or to say that the salt dissolved in water because it is water-soluble is to offer an explanation that references what constitutes salt—namely sodium and

[34] If one is just enforcing *accepted* roles, then an act cannot be said to be discriminatory, as, for example, in the case of a parent–child relationship. There must be a critique of the reproduction of the status to identify something as discriminatory vis-à-vis that status. If, for example, an employer says a woman is being dismissed not because she is pregnant but because she needs physical accommodations to perform her job, that act can be identified as pregnancy discrimination only by referencing what pregnancy consists of (the physical changes entailed in the condition) and by relying on a moral theory designating what's fair given what pregnancy is.

[35] Petri Ylikoski, *Causal and Constitutive Explanation Compared*, 78 ERKENNTNIS 277, 277–78 (2013). Robert Cummins calls the search for constitutive explanation "property theories," which seek to "explain the properties of a system *not* in the sense in which this means 'Why did S acquire P?' or 'What caused S to acquire P? but, rather, 'What is it for S to instantiate P?', or, 'In virtue of what does S have P?'" ROBERT CUMMINS, THE NATURE OF PSYCHOLOGICAL EXPLANATION 14–15 (1983) (emphasis added). Cummins goes on to give the example of the kinetic theory of heat as a property theory because "it explains temperature in a gas by explaining how temperature is instantiated in a gas; it does not, by itself, explain changes in temperature." *Id.*

[36] A constitutive explanation states a relation of dependence (on the property of fragility) that holds constant the triggering conditions (the spoon strike) just like a counterfactual causal explanation—e.g., "the glass broke because it was struck by the spoon"—holds constant the constitutive properties of the glass (its fragility) and imagines varying the triggering condition (strike vs. no strike of the spoon).

[37] This example is taken from Ylikoski, *supra* note 35, at 278–80; *see also* NANCY CARTWRIGHT, HUNTING CAUSES AND USING THEM: APPROACHES IN PHILOSOPHY AND ECONOMICS 14–23 (2007) (discussing the causal capacities of complex nonmodular systems like a carburetor that can be understood only by reference to the system's overall geometry and structure of its component parts).

chloride atoms combined in an ionic bond—to make sense of a dispositional property of the substance when met with a different substance that consists of one atom of oxygen bound to two hydrogen atoms with covalent bonds.[38]

To ask if something happened "because of race" similarly calls for a constitutive explanation, one that references the complex system of social meanings and relations that make up the very category.[39] To say, for example, "I was not given a traffic ticket when I was pulled over for speeding because I am white" is to offer a constitutive explanation that references what constitutes WHITENESS—namely, a social type entailing a presumption of noncriminality and deservedness of leniency and respect. If I further contend that it was racially *discriminatory* that I was not given a ticket, then I am invoking a thick moral claim that is only intelligible if one understands what whiteness consists of. I am saying that the act ought to be condemned because of the manner in which it relies upon meanings or perpetuates understandings that make out the social kind WHITE.

Constitutive explanations proffer counterfactual dependence: to say that a given system has a particular causal capacity or dispositional properties because of how it is constituted means that if you changed the parts and organization of the system, it would have different causal capacities or dispositional properties. But it would also be a different system; the category doing the causing would *be* a different category, just as two different allotropes are constituted by different structural arrangements and bonds of the same atoms. Diamond and graphite are both allotropes of the same atom, but *because of* the different structural arrangements and bonds (i.e., what constitutes them) they are different substances, which, in turn, display different dispositional chemical and physical properties.

That is the heart of the difference between counterfactual causal claims and constitutive causal claims: counterfactual causal claims track an etiological dependence, whereas constitutive causal claims track the

---

[38] As Cummins puts it: "To explain a dispositional regularity, then, we must explain how or why manifestations of the disposition are brought about given the requisite precipitating conditions." CUMMINS, *supra* note 35, at 19. Much of natural science is about offering constitutive explanations. *See, e.g.*, CARL F. CRAVER, EXPLAINING THE BRAIN: MECHANISMS AND THE MOSAIC UNITY OF NEUROSCIENCE 107–12 (2007).

[39] Or consider the analogy to witchcraft offered by Karen Fields and Barbara Fields in their brilliant book *Racecraft*: "Witchcraft . . . acquires perfectly adequate moving parts when a person acts upon the reality of the imagined thing; the real action creates evidence for the imagined thing. . . . In Luther's day, learned jurists and ecclesiastics produced mountains of such evidence." KAREN E. FIELDS & BARBARA J. FIELDS, RACECRAFT: THE SOUL OF INEQUALITY IN AMERICAN LIFE 22 (2012). In an enchanted world where witches are feared entities, the sentence "She was killed because she was a witch" calls forth a constitutive explanation about the social kind "witch," one that refers to the content and structure of a system of religious beliefs and cultural understandings that made it possible to apprehend such a thing as "witch" and how those constitutive elements entail fear and violent rejection of that kind.

dependence of capacities of systems to the precise "properties of parts and/or their organization," which make the system *what it is*.[40] The proposition "If I had been black, I would have gotten a ticket" is really just another way of saying the social type BLACK is treated P-way in X encounters and the social type WHITE is treated G-way in X encounters. The proffered counterfactual does not get us to the discriminatory label without a moral theory that social types BLACK and WHITE both ought to be treated G-way (or P-way) in X encounters. But note that such a moral theory can only be advanced in light of what it is to have stratified racial social types BLACK and WHITE; we can only pick out acts that ought to be condemned in the specific way the label "racially discriminatory" condemns if we know which of the meanings, practices, and relations that constitute social kinds by race we want to disavow. And that, I take it, is the very point of the antidiscrimination project: to transform the social meaning of social categories that have—for so long, in so many domains—been infused with disfavor and disadvantage.

Defining discrimination as a thick ethical concept whose detection demands an explanation of constitutive dependence—namely, analyzing how the distinctive wrongfulness of the action or practice is dependent upon what the category of race consists of—as opposed to identifying a value-neutral fact of counterfactual dependence—namely, determining if the person's (or unit's) racial trait had been different, but nothing else, whether the action would have taken place—is a major conceptual shift. It requires abandoning common ways of talking about race as an attribute or trait and revising the hegemonic way causality is invoked to identify discrimination. Before the reader despairs that I am offering a much too complicated and fussy conceptual apparatus to deal with a fairly obvious set of questions, let me suggest that the possibility of thinking of race as an isolated unit attribute that can be manipulated without diffusing the very meaning of the unit for purposes of detecting discrimination is only possible because of our own "prenotions" about the category from living in a deeply racialized society.[41] If we properly understand the category, then it is impossible to think of it in terms of an isolated treatment.

Sometimes a fantastical analogy shakes away the blinders of our own taken-for-granted categories. Consider an island society where the categories of social stratification are binary: Royal and non-Royal. The privileged class, Royal, wears purple capes and carries sticks. Their cultural tastes define what

---

[40] Ylikoski, *supra* note 35, at 290. To be system S_1 is to have parts and organization $P\text{-}O_1$ and to be system S_2 is to have parts and organization $P\text{-}O_2$ and so on and so forth.

[41] *See* DURKHEIM, *supra* note 17, at 39–46; EMIRBAYER & DESMOND, *supra* note 16, at 31–33, 49 (calling for a "rigorous and methodological delineation of the problem at hand, rather than an uncritical acceptance of definitions already provided by folk wisdom and/or academic culture").

is deemed high and valuable in this society, they occupy more prestigious occupations on the island, hoard more resources, etc.—you get the idea. To say that Royal is a social construct means that Royal is not merely the attribute of wearing purple capes and carrying sticks. Royal is a cultural category of thought and action constituted by a complex set of social relations and meanings that interactionally give import to indices of the category, such that the stick becomes a royal scepter and the cape becomes a sacred robe.

A visiting anthropologist observes that in this society non-Royals step off the sidewalk when Royals are walking on the sidewalk. In order to properly characterize that action in a thick way—as non-Royal debasement (or, conversely, respectful Royal obedience)—she would need to first understand how the categories Royal and non-Royal are constituted in this society by analyzing social relations and cultural meanings. If she were to characterize the action as spontaneous adjustment to scarce sidewalk space or the expression of idiosyncratic sidewalk-versus-road personal preferences, she would misunderstand the *meaning* of the action in the culture. The visiting anthropologist would have no way of making sense of the fact that people with certain attributes (lacking capes and sticks) consistently step off the sidewalk when people with other attributes (capes and sticks) approach, or, more importantly, of making sense of why the former are consistently disadvantaged in other arenas of social and economic life. Nor would she be able to make sense of the moral dimension to the debate in the island's political body over a bill requiring that all people walk on the sidewalk all of the time, especially when other laws on the island defend the right to express personal preferences in the market or intimate affairs, and when the culture generally applauds solving coordination problems without government direction.

The visiting anthropologist could not classify the sidewalk action as non-Royal debasement simply by asking, "If I changed an isolated trait about a person (cape, stick) and *nothing else* changed about that person, would other pedestrians have remained on the sidewalk?" A lot of things change if she makes that manipulation in this society: namely, people no longer perceive a Royal to be walking down the sidewalk! It is only possible for those in the culture to react to the category of Royal because they recognize the indices of Royal to mean something significant beyond the holding of a stick or wearing of a cape. The way those meanings are embodied in the sidewalk behavior is precisely what the anthropologist is trying to capture so that she might properly understand the sidewalk behavior, which means she is seeking a constitutive explanation. She must explain how the structures of social relations and cultural meanings that constitute the very categories of

Royal and non-Royal make this action recognizable as non-Royal debasement, as opposed to spontaneous adjustment to scarce sidewalk space or randomly distributed preferences for when to walk on the sidewalk versus the road.

And yet, in our culture, the most common way of defining discrimination depends upon conceptualizing one of our society's most important categories of social stratification as a trait that can be isolated and manipulated as a treatment, instead of as a social construct whose structure of meaning must be analyzed in order to make sense of a particular instance of action, patterned practice, or policy.

II. THE COUNTERFACTUAL CAUSAL MODEL OF DISCRIMINATION IN SOCIAL SCIENCE AND LAW

This is not the doctrinal Part. Rather, this Part reveals that a common approach—perhaps the most common approach—in both law and social science to detecting discrimination is to measure the treatment effect of race in the counterfactual sense. Part III carefully explicates this counterfactual causal model (CCM), but for purposes of this Part the succinct sketch presented in the Introduction is sufficient to identify its pervasiveness, either explicitly or implicitly, in both domains.

I provide an overview of both law and social science because my sense is that the hegemony of the CCM is the product of multiple, interactive developments including, on the one hand, an intentional drive to narrow the ambits of equal protection doctrine, and on the other hand, quantitative methods for detecting discrimination filling in a substantive meaning of the concept DISCRIMINATION that is otherwise lacking with indeterminate formal formulations such as "equal protection," "treating like cases alike," or "similarly situated."[42] At the risk of being repetitive, this Part does not evaluate whether the studies and cases cited in this Section *correctly* detect discrimination, because my argument is that we can never answer that question without a prior sociological account of the category and moral theory of fair decision-making or resource allocation in the relevant domain in light of what the category is in a particular place and time. My aim is simply to show that these studies often proceed *as if* the aim of the exercise is to get at a pure treatment effect of race or ethnicity presuming there is an objective trait there to be gotten at after stripping away confounders.

Many social scientists explicitly embrace a definition of discrimination as the causal effect of race (or ethnicity or sex). For example, the prestigious

[42] Explaining how and why this model came to occupy the position it does is a project for another article altogether.

National Research Council's report *Measuring Racial Discrimination* says that "to measure discrimination researchers must answer the counterfactual question: What would have happened to a nonwhite individual if he or she had been white?"[43] The report goes on to illustrate both the logic of counterfactual causal inference and the conceptualization of race implicated by the model by appealing to Dr. Seuss's book *The Sneetches*. In that book, Sneetches stamped with a star get access to all sorts of social goods (like killer hot dogs) while starless Sneetches are excluded, get lesser goods (like tofu-dogs), and are despised. The authors note that because we cannot "stamp" race on individuals and thereby directly observe counterfactual outcomes on the same humans with different racial stamps at different times, we must resort to second-best methods for drawing causal inferences about the causal impact of race.[44] Numerous scholars praise the precision of the counterfactual causal definition of discrimination and name audit studies as the gold standard in causal inference.[45] Many notable studies of discrimination in various arenas explicitly or implicitly adopt this model, either by defining discrimination as the causal effect of race alone, or simply by adopting a method that attempts to compare outcomes between units of different racial status that are similarly situated with respect to an exhaustive

[43] NATIONAL RESEARCH COUNCIL, MEASURING RACIAL DISCRIMINATION 77 (Rebecca M. Blank et al. eds., 2004). Indeed, the chapter is entitled "Causal Inference and the Assessment of Racial Discrimination."

[44] Observational studies are placed lower down in the "hierarchy of approaches to data collection." *Id.* at 81.

[45] *See, e.g.*, Sonja B. Starr, *Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police*, 2016 U. CHI. LEGAL F. 485, 487 (2016) ("Auditing has not been tried or even discussed in the law enforcement field, which is surprising because for decades it has been a central tool in antidiscrimination research and civil rights enforcement more generally. It presents safety, legality, and efficacy concerns when applied to policing, but with careful design I argue that these concerns can be overcome. If so, auditing could provide something observational research usually cannot: causally rigorous analysis of police discrimination in a real-world setting."); Lincoln Quillian, Book Review, 35 CONTEMP. SOC. 88, 89 (2006) (reviewing NATIONAL RESEARCH COUNCIL, *supra* note 43) ("Racial discrimination is defined with reference to counterfactual notions of causality developed in statistics: racial discrimination for an individual is the difference in an outcome if an individual were of one race contrasted to another race. This definition is cleverly illustrated with reference to the Dr. Seuss story *The Sneetches*. The fundamental methodological problem of measuring discrimination results because, unlike the Sneetches, we do not observe outcomes for each person under both racial conditions. Instead, we must use indirect techniques to estimate the magnitude of racial discrimination."). A more recent National Academies report also endorses the counterfactual causal definition of discrimination, with the qualification that laboratory methods for detecting it face external validity challenges. NATIONAL ACADEMIES OF SCIENCES, PROACTIVE POLICING: EFFECTS ON CRIME AND COMMUNITIES 256 (2018) ("Studies of behavior in a simulated laboratory environment[] offer the benefit of studying how people make decisions in situations where, by construction, the only variable that differs across encounters is the race of the subject.").

set of imaginably decision-relevant characteristics (often limited by what is available in administrative data).[46]

Nothing in the text of the Constitution necessitates specifying racial discrimination as the CCM; the relevant text merely articulates practically indeterminate formal principles. The Fourteenth Amendment commands that no state shall "deny to any person within its jurisdiction the equal protection of the laws"; here, as in any other principle of formal equality, there is no way to *apply* the principle without an independent account of what equal protection means and the values underlying such a judgment.[47]

The Supreme Court has stated that "[t]he Equal Protection Clause . . . is essentially a direction that all persons similarly situated should be treated alike."[48] Following the Supreme Court's decision in *Washington v. Davis*,[49]

---

[46] *See, e.g.*, Joseph G. Altonji & Rebecca M. Blank, *Race and Gender in the Labor Market*, *in* HANDBOOK OF LABOR ECONOMICS 3143, 3192 (1999), https://ideas.repec.org/h/eee/labchp/3-48.html [https://perma.cc/MF32-MF8G] ("To investigate the presence of discrimination, one would like to be able to compare the outcomes of individuals in the same job who are identical in all respects that are relevant to performance but who differ only in race, ethnicity or gender."); Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANN. REV. SOC. 181, 184 (2008) ("While statistical models represent an extremely important approach to the study of race differentials, researchers should use caution in making causal interpretations of the indirect measures of discrimination derived from residual estimates."); Lincoln Quillian, *New Approaches to Understanding Racial Prejudice and Discrimination*, 32 ANN. REV. SOC. 299, 302 (2006) ("To estimate the magnitude of discrimination in a particular context then involves answering a counterfactual question: What would the treatment of target group members have been if they had been dominant group members? This counterfactual notion of discrimination measurement corresponds to the use of counterfactuals in the causal effects literature (Winship & Morgan 1999). Discrimination is the causal effect of race on an outcome with other factors held constant."); Roland G. Fryer, Jr., *An Empirical Analysis of Racial Differences in Police Use of Force* 35–38 (Nat'l Bureau of Econ. Research, Working Paper No. 22399, 2018).

[47] U.S. CONST. amend. XIV, § 1. The entire "equality of what" debate in egalitarianism is about this same issue. *See, e.g.*, G.A. Cohen, *Equality of What? On Welfare, Goods and Capabilities*, 56 LOUVAIN ECON. REV. 357, 357 (1990); Ronald Dworkin, *What Is Equality? Part 1: Equality of Welfare*, 10 PHIL. & PUB. AFF. 185, 185 (1981); Ronald Dworkin, *What Is Equality? Part 2: Equality of Resources*, 10 PHIL. & PUB. AFF. 283, 283 (1981); Amartya Sen, *Equality of What?*, *in* EQUAL FREEDOM: SELECTED TANNER LECTURES ON HUMAN VALUES 307, 307 (Stephen Darwall ed., 1995). And for a brilliant critique of the terms in which this debate is framed, see Elizabeth S. Anderson, *What Is the Point of Equality?*, 109 ETHICS 287, 287 (1999). Similar arguments about the need for a substantive account of the values behind equality or the material to be equalized have been artfully developed in other subject areas such as tort law, *see* Jules Coleman & Arthur Ripstein, *Mischief and Misfortune (Annual McGill Lecture in Jurisprudence and Public Policy)*, 41 MCGILL L.J. 91, 91 (1995), and of course extensively in antidiscrimination law, *see* Larry Alexander, *What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies*, 141 U. PA. L. REV. 149, 151 (1992); Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination*, 58 U. MIAMI L. REV. 9, 9 (2003); Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 107–08 (1976); Post, *supra* note 30, at 14.

[48] City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432, 439 (1985).

[49] 426 U.S. 229, 242 (1976) ("[W]e have not held that a law, neutral on its face and serving ends otherwise within the power of government to pursue, is invalid under the Equal Protection Clause simply

the formal equal protection demand to treat like cases alike has boiled down to three doctrinal formulations of discrimination under the Equal Protection Clause in federal courts: (1) explicit racial classification;[50] (2) facially neutral law or policy that has a disparate racial impact and was motivated by discriminatory animus or intent;[51] and (3) facially neutral law or policy that is applied in an intentionally discriminatory manner and results in a disparate impact.[52]

The practical result is that in order for a litigant to show that a particular practice that does not facially classify by race was discriminatory, he or she must show that the practice had discriminatory effects and that the decision-maker intended to discriminate. Circularity at its finest: Discrimination = discriminatory effect + discriminatory intent.[53] The term we are trying to define enters into the terms doing the definitional work and is seemingly essential to it, as otherwise we have no way of distinguishing permissible from impermissible effects and intents.

Faced with an indeterminate circular definition, one can see the attraction of the CCM: it seems to provide a way of distinguishing between discriminatory and nondiscriminatory effects by reference to what is essentially functionally rational criteria. In selective prosecution and enforcement cases (i.e., discrimination by law enforcement in selection of targets, or discrimination by prosecutors regarding cases referred by law enforcement, respectively), that is usually articulated as the "similarly situated" test: "[C]ourts have required a defendant to make a credible showing that a similarly situated individual of another race or ethnicity could

---

because it may affect a greater proportion of one race than of another. Disproportionate impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution. Standing alone, it does not trigger the rule that racial classifications are to be subjected to the strictest scrutiny and are justifiable only by the weightiest of considerations." (internal citations omitted)).

[50] *See, e.g.*, Brown v. Bd. of Educ., 347 U.S. 483, 493 (1954).

[51] *See, e.g.*, Hunter v. Underwood, 471 U.S. 222, 233 (1985) (holding an Alabama law disenfranchising those convicted of certain crimes invalid because it was enacted with a racially discriminatory purpose and had a racially disparate impact).

[52] *See, e.g.*, Yick Wo v. Hopkins, 118 U.S. 356, 373–74 (1886) ("Though the law itself be fair on its face and impartial in appearance, yet, if it is applied and administered by public authority with an evil eye and an unequal hand, so as practically to make unjust and illegal discriminations between persons in similar circumstances, material to their rights, the denial of equal justice is still within the prohibition of the Constitution."); *see also* Pyke v. Cuomo, 258 F.3d 107, 110 (2d Cir. 2001) ("[A] plaintiff seeking to establish a violation of equal protection by intentional discrimination may proceed in 'several ways,' including by pointing to a law that expressly classifies on the basis of race, a facially neutral law or policy that has been applied in an unlawfully discriminatory manner, or a facially neutral policy that has an adverse effect and that was motivated by discriminatory animus.").

[53] For example, consider the following explication: "[A] plaintiff alleging the discriminatory application of a neutral law or policy must demonstrate that the application of the policy was motivated by discrimination." Ali v. Connick, 136 F. Supp. 3d 270, 279 (E.D.N.Y. 2015).

have been subjected to the same law enforcement action as the defendant, but was not."[54]

The obvious problem with defining unlawful discrimination under the Fourteenth Amendment as synonymous with functional irrationality is that it undercuts the entire logic of equal protection jurisprudence's graded scrutiny scale under which allegations of racial discrimination are subject to the most exacting demands for rationality.[55] Some prior principle is necessary to explain why the state need only proffer some minimally logical account for how a nonracial classification—such as "bunioned," i.e., the status of having a bunion—advances a legitimate state interest, whereas racial classifications require "the most exact connection between justification and classification," in order to "satisfy this searching standard of review," in which the state must demonstrate that the classification was "'narrowly tailored' to achieve a 'compelling' government interest."[56] An obvious candidate would be the famous *Carolene Products*-footnote-four-type answer: that being in the group designated "black" in America is associated with political vulnerability, a history of social exclusion, economic disadvantage, and cultural prejudice, whereas being in the group "bunioned" is not.[57] But once one accepts such a principle to justify the graded scrutiny scale (which, alternatively, could just as well be expressed in terms of what sorts of rationales or state interests will be recognized as legitimate for different types of groups), then the most concerning forms of discrimination cannot be defined as decision-making on the basis of

---

[54] United States v. Duque–Nava, 315 F. Supp. 2d 1144, 1153 (D. Kan. 2004).

> This element of a "similarly situated individual" has been applied in claims of selective prosecution, to require a defendant to show not only that his racial or ethnic group is prosecuted more than another group, but that a similarly situated individual in another group was not prosecuted for the same offense.

Id.

[55] "It should be noted, to begin with, that all legal restrictions which curtail the civil rights of a single racial group are immediately suspect. That is not to say that all such restrictions are unconstitutional. It is to say that courts must subject them to the most rigid scrutiny." Korematsu v. United States, 323 U.S. 214, 216 (1944).

[56] Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) (internal citations omitted).

[57] The famous footnote emerged from a case where the Court upheld a statute forbidding the interstate commerce in "filled milk," stating that "the existence of facts supporting the legislative judgment is to be presumed, for regulatory legislation affecting ordinary commercial transactions"; however, "[t]here may be narrower scope for operation of the presumption of constitutionality when legislation appears on its face to be within a specific prohibition of the Constitution," including "statutes directed at particular religious . . . or racial minorities, . . . [or] against discrete and insular minorities . . . ." United States v. Carolene Prods. Co., 304 U.S. 144, 152–53 (1938).

"irrelevant" factors.[58] And yet the entire logic of the CCM is to define discrimination as a form of irrationality, which requires units (persons, neighborhoods, etc.) to present with identical functionally relevant characteristics in order to elicit differential effects.

Many (but not all) courts have operationalized both requirements of an equal protection claim—differential outcomes by protected group (discriminatory effect) and differential outcome because of the protected group status (discriminatory intent)—in counterfactual terms.[59] As the examples in the remainder of this Section show, litigants faced with this evidentiary burden often turn to quantitative methods to attempt to show both that there are differential outcomes by group and that, if those groups are otherwise similarly situated with respect to all other relevant characteristics or traits, race was the "cause" by elimination. To pursue this strategy, a litigant needs data on the relevant units indicating raced status (either at the individual or aggregate level depending on the unit of analysis), the outcome of interest (e.g., pedestrian or traffic stop, adverse police action, arrest, etc.), and other variables theoretically germane to the outcome. Armed with this data, a litigant then just needs to hire a statistical expert to use some methodologically sophisticated techniques to try to demonstrate that differential outcomes persist between "similarly situated" units.

The promise of quantitative evidence (or experimental or quasi-experimental evidence) is obvious: It offers a way to proceed despite the doctrinal narrowing of the ambits of equal protection initiated by *Washington*

---

[58] The so-called "colorblindness" approach has often sounded in this "irrelevance" language. *See, e.g.*, *Parents Involved in Cmty. Sch.*, 551 U.S. at 730 ("Allowing racial balancing as a compelling end in itself would 'effectively assur[e] that race will always be relevant in American life, and that the "ultimate goal" of "eliminating entirely from governmental decisionmaking such irrelevant factors as a human being's race" will never be achieved.'" (internal citations omitted)). Presumably "irrelevance" here is meant aspirationally, because the statuses about which we are the most concerned will be the objects of discrimination are the ones where the status marker is highly correlated with other important social indicators—which of course is precisely *why* we are concerned about those statuses being the object of discrimination in the first instance—and therefore, it is empirically inaccurate to call the statuses irrelevant.

[59] This Section presents only a few examples of common formulations when the plaintiff is alleging an equal protection violation of "a facially neutral" law or policy. In a case challenging, on First and Fifth Amendment grounds, the "passive enforcement" practice of the Selective Service in which they initiated prosecutions only against those individuals who either self-reported refusal to register or were reported by others, the Supreme Court said, "It is appropriate to judge selective prosecution claims according to ordinary equal protection standards. Under our prior cases, these standards require petitioner to show both that the passive enforcement system had a discriminatory effect and that it was motivated by a discriminatory purpose." Wayte v. United States, 470 U.S. 598, 608 (1985) (internal citation omitted); *see* Chavez v. Ill. State Police, 251 F.3d 612, 635–36 (7th Cir. 2001) ("To show a violation of the Equal Protection Clause, plaintiffs must prove that the defendants' actions had a discriminatory effect and were motivated by a discriminatory purpose.").

v. Davis.[60] Some litigants can show that differential outcomes persist after controlling for every conceivable decision-relevant variable. But many cannot do so. The challenge for litigants to demonstrate "similarly situated" units is most pronounced when the historical forces of racial and ethnic group formation have created separate social and physical worlds for different groups.[61] Defendants can, drawing on their knowledge about how race and ethnicity structure the social world, easily construct a post hoc list of variables that are unequally distributed by race or ethnicity (making it very difficult, if not impossible, to find the applicable counterfactual) that could theoretically "justify" the disparate treatment.

Quantitative evidence based on multivariable regressions can work well for litigants in situations where the "effect of race" on an outcome—meaning the statistical significance of the variable measuring race (or racial

---

[60] As many have noted, judges politically dedicated to narrowing the uses of equal protection challenges have actively done so under many doctrinal formulations over recent decades. This is a massive topic but see, for example, IAN HANEY LÓPEZ, WHITE BY LAW: THE LEGAL CONSTRUCTION OF RACE 158–62 (2006); Mario L. Barnes et al., *A Post-Race Equal Protection?*, 98 GEO. L.J. 967–1004 (2010); Neil Gotanda, *A Critique of "Our Constitution Is Color-Blind,"* 44 STAN. L. REV. 1, 2–62 (1991); Reva B. Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278–367 (2011); and Kimberlé Williams Crenshaw, *Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law*, 101 HARV. L. REV. 1331–87 (1988).

[61] Sometimes, but not always, courts recognize this dilemma. In the *Pyke* case, which did not involve quantitative evidence, Native American opponents of gambling living on a reservation claimed an equal protection violation alleging that the state of New York failed to provide them with adequate police protection "because the persons in need of protection were Native Americans," but the defendants' position was that they did not send the state police "to intervene in the events occurring on the reservation because the Indian tribe exercises a considerable measure of self-governance on the reservation, and because the violence on the reservation threatened the safety of the state police officers." Pyke v. Cuomo, 258 F.3d 107, 108 (2d Cir. 2001). The district court granted summary judgment to the defendants, finding that the plaintiffs failed to allege an express racial classification or show that other similarly situated groups were treated differently. *Id.* The Second Circuit reversed, holding that in a police discrimination case (as opposed to selective prosecution), as long as plaintiffs "allege and establish that the defendants discriminatorily refused to provide police protection because the plaintiffs are Native American, plaintiffs need not allege or establish the disparate treatment of otherwise similarly situated non-Native American individuals." *Id.* at 109. The court recognized that

> [i]t would be difficult, if not impossible, to find other individuals whose situation is similar to Native Americans living on a reservation and exercising a substantial measure of self-government independent of New York State. Plaintiffs would probably be incapable of showing similarly situated individuals who were treated differently. If the rule were as framed by the district court, police authorities could lawfully ignore the need of Native Americans for police protection on the basis of discriminatory anti-Indian animus.

Id. Even if litigants are not required to come forward with similarly situated units that were treated differently, many courts interpret legal rules such as the following to mean the status caused the outcome in the counterfactual sense: "Once the plaintiff shows that the application was so motivated, 'at least in part,' the defendant must show that the same result would have occurred even without consideration of the plaintiffs[']race or national origin." Ali v. Connick, 136 F. Supp. 3d 270, 279 (E.D.N.Y. 2015) (citing United States v. City of Yonkers, 96 F.3d 600, 612 (2d Cir. 1996)).

composition if the unit is an aggregation)—survives inclusion of a range of variables plausibly relevant to the outcome of interest.[62] In *Floyd v. City of New York*, for instance, the plaintiffs demonstrated that the racially disparate impact of the New York Police Department's stop-and-frisk (SQF) policy could not be explained by "legitimate" bases for police decisions, such as high crime rates in minority neighborhoods.[63] The exercise of making units similarly situated included variables such as violent crime complaint rates, which the police may have actually consulted in making enforcement allocations.[64] But the models submitted by the plaintiff's expert did not stop there. They also included other variables—such as unemployment and housing vacancy rates—that *could* be theoretically relevant to the number of stops, but it is doubtful that the NYPD actually consulted these variables in its decision-making.[65] The logic was that if racial composition remained a statistically significant predictor of the intensity of stop-and-frisks even after controlling for every other possible variable, then it must represent the residual of racially discriminatory intent. The *Floyd* plaintiffs were successful in demonstrating that the correlation between minority composition (at the precinct or census-tract level) and SQF rates survived the inclusion of a bevy of other variables plausibly relevant to police allocation or enforcement decisions with lots of little stars of statistical

---

[62] I put "effect of race" in scare quotes to indicate that the term should be understood as "statements of association between the racial classification and a predictor or explanatory variable across individuals in a population." Zuberi, *supra* note 31, at 178.

[63] 959 F. Supp. 2d 540, 589 (S.D.N.Y. 2013) (accepting statistical findings that "the NYPD carries out more stops in areas with more black and Hispanic residents, even when other relevant variables are held constant. The best predictor for the rate of stops in a geographic unit—be it precinct or census tract—is the racial composition of that unit rather than the known crime rate").

[64] Interestingly, the City's expert rebuttal report seemed to take issue even with the model that serious felony crime complaint rates should predict SQF rates because such a model fails to "confront the historic shift at NYPD away from a primary mission of responding to crime to a mission of preventing crime through proactive and crime targeted police vigilance." Report of Dennis C. Smith, Ph.D., at 4, Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013) (No. 08 Civ. 1034 (SAS)). Besides the befuddling sentences, such as "[a]ny credible analysis of the determinates of stop and frisk activity must first control for the impact of evidence-based management practices before trying to parse out any other factors that may or may not have contributed to stop and frisk patterns," it is left entirely unspecified in the defendant's report what indices the police *would* look at to make enforcement decisions in this new era of scientifically managed crime prevention, if not recent past crime complaints. *Id.* at 5.

[65] For example, the plaintiff's expert analysis of SQF rates by spatial aggregations (police precincts in the first expert report, Report of Jeffrey Fagan, Ph.D., at 7–8, 12, *Floyd*, 959 F. Supp. 2d 540 (No. 08 Civ. 1034 (SAS)), or census tracks in the second supplemental report, Second Supplemental Report of Jeffrey Fagan, Ph.D., at 16–17, *Floyd*, 959 F. Supp. 2d 540 (No. 08 Civ. 1034 (SAS))) included variables about which there appeared to be little to no evidence that police actually did take into direct account in making patrol strength or tactics allocations, such as unemployment, median household income, housing vacancy, or residential mobility. Report of Jeffrey Fagan, Ph.D., *supra*, at 31–32.

significance.[66] But the same analysis might not succeed in a hypersegregated city where the particularities of its racial history have produced a social geography with few if any majority-black neighborhoods that share all relevant characteristics with majority-white neighborhoods.

Consider, for example, *United States v. Jones*, a selective prosecution challenge to Project Exile, which diverted gun cases from state to federal court, where defendants faced significantly enhanced prison sentences.[67] About 90% of the defendants in this program were black, and cases were diverted from just two jurisdictions: the cities of Norfolk and Richmond, Virginia, which had a jury pool that was about 75% African-American, whereas other cases were diverted to the Richmond Division of the Eastern District of Virginia, which had a jury pool that was about 10% African-American; the plaintiffs submitted evidence that an AUSA "stated that one goal of Project Exile is to avoid 'Richmond juries.'"[68] Nonetheless, the court stated that, while

> Project Exile would be vulnerable on selective prosecution grounds if African-American defendants were routinely diverted from state to federal prosecution while prosecutors allowed similarly situated Caucasian defendants to remain in state court[,] . . . [plaintiff] presents no evidence of Caucasian defendants similarly situated to defendant Jones evading diversion to federal court.[69]

The court was moved by the Government's claim that "[t]hose who implemented Project Exile have targeted cities in which violent crime is most

---

[66] "After controlling for crime (prior month) and other tract social and economic characteristics plus patrol strength, the percent black or Hispanic in the census tract significantly and positively predicts the likelihood of Black suspects or Hispanic[] suspects being stopped relative to Whites." Second Supplemental Report of Jeffrey Fagan, Ph.D., *supra* note 65, at 19.

[67] "Under Project Exile, local police review each firearm-related offense to determine whether the conduct alleged also constitutes a federal crime. *See, e.g.*, 18 U.S.C. §§ 922(g) and 924(g) (prohibiting the possession of firearms by certain persons)." United States v. Jones, 36 F. Supp. 2d 304, 307 (E.D. Va. 1999). Establishing discriminatory effect and discriminatory intent is further complicated by the deference courts give certain categories of officials. Although the Supreme Court has instructed courts to "judge selective prosecution claims according to ordinary equal protection standards," Wayte v. United States, 470 U.S. 598, 608 (1985), it has also gone to great lengths to say that separation of powers requires courts to respect the discretion of prosecutors in selecting targets for prosecution. *See* United States v. Armstrong, 517 U.S. 456, 464 (1996) ("A selective-prosecution claim asks a court to exercise judicial power over a 'special province' of the Executive." (quoting Heckler v. Chaney, 470 U.S. 821, 832 (1985))); United States v. Davis, 793 F.3d 712, 720 (7th Cir. 2015) ("*Armstrong* was about prosecutorial discretion. The defendants assumed that state and federal law-enforcement agents arrested all those they found dealing in crack cocaine, and they suspected that the federal prosecutor was charging the black suspects while letting the white suspects go. The Supreme Court replied that federal prosecutors deserve a strong presumption of honest and constitutional behavior, which cannot be overcome simply by a racial disproportion in the outcome, for disparate impact differs from discriminatory intent.").

[68] *Jones*, 36 F. Supp. 2d at 307–08.

[69] *Id.* at 311.

prevalent," which, because of the history of racial segregation and discrimination, in most American cities are almost always the same spaces that are predominantly African-American.[70]

Or consider how the counterfactual model of discrimination plays out in a series of "phony stash house" cases from across the country, in which various law enforcement agencies use undercover agents or confidential informants recruit targets to participate in an armed robbery of a nonexistent stash house.[71] Once the robbery is arranged and coconspirators are recruited, the targets are arrested and charged. Defendants have sought to dismiss indictments on grounds of racial profiling or sought discovery in the hopes of establishing as much. Most—though not all[72]—judges have denied defendants' motions on the grounds that they could not prove that "similarly situated" whites were not targeted.[73] In the face of overwhelming disparate

---

[70] *Id.* at 312.

[71] "Since 2006, the Bureau of Alcohol, Tobacco, Firearms, and Explosives (the 'ATF') has engaged in sting operations wherein undercover agents present individuals in this District with an opportunity to rob a fictitious drug stash house." United States v. Brown, 299 F. Supp. 3d 976, 983 (N.D. Ill. 2018).

> Developed by the ATF in the 1980s to combat a rise in professional robbery crews targeting stash houses, reverse sting operations have grown increasingly controversial over the years, even as they have grown safer and more refined. For one, they empower law enforcement to craft offenses out of whole cloth, often corresponding to statutory offense thresholds. Here, the entirely fictitious 10 kilograms of cocaine triggered a very real 20-year mandatory minimum for Washington, contributing to a total sentence of 264 months in prison—far more than even the ringleader of the conspiracy received. For another, and as Washington claimed on multiple occasions before the District Court—and now again on appeal—people of color are allegedly swept up in the stings in disproportionate numbers.

United States v. Washington, 869 F.3d 193, 197 (3d Cir. 2017), *cert. denied*, 138 S. Ct. 713 (2018); John Diedrich & Raquel Rutledge, *ATF Uses Rogue Tactics in Storefront Stings Across Nation*, MILWAUKEE J. SENTINEL (Dec. 7, 2013), http://archive.jsonline.com/watchdog/watchdogreports/atf-uses-rogue-tactics-in-storefront-stings-across-the-nation-b99146765z1-234916641.html [https://perma.cc/W8U3-264U].

[72] *See, e.g.*, *Davis*, 793 F.3d at 715, 723 (holding that limited discovery is appropriate in face of statistics that of twenty stash house stings, seventy-five defendants were African-American, thirteen were Hispanic, and only six were non-Hispanic whites).

[73] *See, e.g.*, United States v. Payne, No. 12 CR 854 (CRN), slip. op. at 3 (N.D. Ill. Jan. 20, 2015) (where statistics showing that only one of twenty-six stash house cases had a white defendant and almost 80% of defendants were African-American were insufficient to demonstrate discriminatory effect because there is "no evidence . . . of other similarly situated individuals of a different race who were not prosecuted"). In *Lamar*, the judge rejected the defendant's claim that, given the existence of white individuals with a similar criminal record, this pattern constituted sufficient evidence of discriminatory effect to compel discovery on how the DEA selection took place:

> The fact that none of the 95 defendants in these cases is white does not constitute "some evidence" of either discriminatory effect or discriminatory intent. A close review of the complaints in these eighteen cases reveals that 76 of the 95 defendants in these cases were recruited not by a DEA informant, but instead by a coconspirator.

United States v. Lamar, No. 14 CR 726 (PGG), 2015 WL 4720282, at *9 (S.D.N.Y. Aug. 7, 2015). The court went on to reason in counterfactual terms, noting that the defendants conceded:

racial impact, litigants are still expected to show that the effect of race and race alone can be isolated from other factors to support a counterfactual causal account of discrimination. One judge explained that defendants did not meet their burden even to obtain discovery where zero of ninety-five defendants in eighteen robbery sting cases in the Southern District of New York were white because, among other things, they did not point to incredibly specific evidence that "any white person who had been sentenced under the Robbery Guidelines, or who had served time in a New York state prison for a violent felony, ever told a DEA informant about past involvement in robbing drug dealers but was not targeted for a sting operation."[74]

Even if discovery is provided, it is not difficult for law enforcement to come forward with plausible bases to distinguish potential targets as not "similarly situated," especially post hoc and especially when there are highly unequal distributions between groups of variables that are plausibly rational for law enforcement to consider, such as residence in high crime neighborhoods or criminal history. Consider how this played out in a series of high profile phony stash house cases in Chicago, where the court rejected the defendants' motions to dismiss indictments on equal protection grounds after substantial discovery was provided and extensive warring expert reports submitted, despite the judge's statement that "[i]t is time for these false stash house cases to end and be relegated to the dark corridors of our past."[75] Both experts—for the defendants seeking to establish discrimination and for the government seeking to refute it—made their case in explicitly counterfactual terms by arguing that they could (or could not) isolate the effect of race in enforcement target selection;[76] and the court approached the

---

> [They] have no "direct evidence" that (1) "an informant ever told [the DEA agents involved in this case] about a white person who had robbed drug dealers or who might be interested in robbing drug dealers, and that [the DEA agents involved in this case] indicated that they were [not] interested in pursuing that person"; and (2) the DEA agents involved in this case "instructed informants that the DEA was only interested in targeting non-white people who robbed drug dealers or might be interested in robbing drug dealers."

Id. at *10. Disclosure: I submitted an expert report in this case.

[74] *Lamar*, 2015 WL 4720282, at *15.

[75] *Brown*, 299 F. Supp. 3d at 983–84.

[76] *See, e.g.*, Report of Jeffrey Fagan, Ph.D., at 27–28, *Brown*, 299 F. Supp. 3d 976 (No. 12-CR-632 (RC)) (citing audit studies as the ideal experiment to isolate the causal effect of race and adopting propensity score matching to "simulate[] random assignment to a treatment group–*race*–by matching persons on numerous predictors of treatment assignment"); Expert Report of Max M. Schanzenbach at 12, *Brown*, 299 F. Supp. 3d 976 (No. 12-CR-632 (RC)) (arguing that to interpret residual differences in likelihood of being targeted to discrimination methods must eliminate any systematic differences between groups—including "willing[ness] to participate in a stash house robbery"—except for the racial/ethnic status: "The control variables must capture the underlying differences between black, Hispanic, and white offenders in the sample to such an extent that we can interpret the remaining differences between these

evaluation of the evidence in counterfactual terms.[77] After an in-depth review of the voluminous reports, the court held that the defendants failed to establish either discriminatory effect or intent for many varied and complex reasons; relevant to the discussion here, it concluded that the defendants did not proffer a sufficiently "similarly situated" group of potential white targets to prove race was the causal factor in enforcement selection.[78] The counterfactual causal model proceeds as if race consists in only its signifiers—say an attribute A—and discrimination is defined as the effect of that A on an outcome when everything else—call them all of the other Xs—are held constant. If there is no prior sociological account of the distribution and meaning of the Xs by different racial/ethnic groups and no prior moral equality-of-what theory designating in what fairness consists of *given* the differential distribution and meaning of Xs by group, then there is no limiting principle on what should or should not be stripped away in order to get at some imagined solid state of race or ethnicity.

This logic is also on display in the use of racial-specific or ethnic-specific crime rates.[79] It turns out that far from the soaring rhetoric that "the Government must treat citizens 'as *individuals*, not "as simply components

---

groups as caused by discrimination."). Sometimes the experts conflate two quite distinct counterfactual questions, namely (1) What would the racial composition of phony stash house sting defendants be if the ATF had been presented with a given quantum of potential black, Hispanic, and white targets that were identical in all Xs (all conceivably relevant variables) except racial/ethnic status?, with (2) What would the racial composition of the phony stash house sting defendants be if the ATF selected enforcement targets in a nondiscriminatory fashion? *See, e.g.*, Expert Report of Max M. Schanzenbach, *supra*, at 4. The former is the CCM that I critique throughout on the grounds that it cannot identify discrimination without a sociological account of the distribution and meaning of Xs by different racial/ethnic group and a moral account of what enforcement target selection processes are fair or just *given* the differential distribution and meaning of Xs by group. The latter proffers a counterfactual normative criterion that is undefined, which would need to be given some content to know what its operation would generate empirically.

[77] [I]f Defendants prove that law enforcement agents would not have pursued these investigations had they been white, dismissal of the charges is warranted. "If not, there would not be a basis to attribute this prosecution to the defendants' race," and the case must proceed to a resolution of the charges.

Brown, 299 F. Supp. 3d at 995, 1006 (internal citation omitted).

[78] *Id.* at 1013, 1022.

[79] *See, e.g.*, United States v. Duque–Nava, 315 F. Supp. 2d 1144, 1160 (D. Kan. 2004) (discussing Hispanic profiling in Kansas, "the proper consideration here is whether there are any differences in the incidence of traffic violations by different racial and ethnic groups when one considers all such violations, and when one considers the host of reasons officers rely on in effecting a traffic stop"). I do not believe that "similarly situated" with respect to functionally relevant characteristics is really doing the work it claims to be doing in these cases, because the entire basis of concern about a group being the target of discrimination is that one group is systematically different from another in socially important ways. Upon closer inspection, the similarly situated inquiry is precisely where courts—without any explicit acknowledgement—engage a substantive equality-of-what theory, specifying *which* facts that construct racial and ethnic groups will be countenanced and which will not.

of a racial, religious, sexual or national class,"'" courts end up engaging social facts about an individual's ascriptive group when it comes to this type of inquiry.[80] The claim that the black stop rate should be proportional to the rate at which complainants of violent crime describe the perpetrator as black, or that the racial composition of phony stash house sting defendants should match the composition of those arrested for home invasions with a firearm or with "willingness" to participate in a stash house robbery, means that an individual's risk of stop or targeting is evaluated as fair by reference to a proffered empirical fact about their ascribed group (or the "propensity" of their ascribed group).[81] Let's leave aside the difficulty of knowing the true rates at which people ascribed to different ethnic or racial groups engage in the offenses at issue in a particular discriminatory enforcement claim, or the difficulty of assigning people to designated groups in a way that is not endogenous to identifying an instance of the offense.[82] Proposing evidence of differential racial or ethnic group offense rates to defeat a claim of discrimination in a particular enforcement policy only works if one accepts some prior independent moral justification of the policy or practice *given* the claimed racial- or ethnic-specific offense rates. Declaring that the enforcement action happens because of the underlying offending conduct—and not because of race or ethnicity—is simply question-begging. The entire premise of the discriminatory effects inquiry is to designate criteria by which a policy that does not on its face classify by race or ethnicity produces patterns that will be recognized as DISCRIMINATORY.

Many courts take the identical information—proffered racial- or ethnic-specific conduct rates—and insist that it is essential contextual evidence for deciding if the effects really are discriminatory, but *irrelevant* for

---

[80] Metro Broadcasting, Inc. v. FCC, 497 U.S. 547, 602 (1990) (O'Connor, J., dissenting).

[81] Floyd v. City of New York, 959 F. Supp. 2d 540, 584 (S.D.N.Y. 2013) ("The City's experts . . . used a benchmark consisting of the rates at which various races appear in suspect descriptions from crime victims—in other words, 'suspect race description data' . . . [and] assumed that if officers' stop decisions were racially unbiased, then the racial distribution of stopped pedestrians would be the same as the racial distribution of the criminal suspects in the area."); *see also Brown*, 299 F. Supp. 3d at 1012–13 (discussing the Government's expert report, "[a]s a 'counterfactual' to Professor Fagan's findings, Professor Schanzenbach . . . found that black, white, and Hispanic defendants in the stash house cases had statistically similar criminal histories, and that on average, black defendants had more convictions, arrests, and sentences, and were more likely to have been convicted or arrested for a weapons offense than white defendants[;] . . . [o]f those arrested for weapons offenses, 72.5 percent were black and 15 percent were Hispanic[;] . . . for home invasion with a firearm, 75.4 percent were black and 12 percent were Hispanic").

[82] Aliya Saperstein & Andrew M. Penner, *Racial Fluidity and Inequality in the United States*, 118 AM. J. SOC. 676, 712 (2012) (showing that external racial identification changes over time by redefining successful or high-status people as white (or not black) and unsuccessful or low-status people as black (or not white)).

determining whether to infer discriminatory intent.[83] Consider, for example, that many courts ruled that public perception of crack as a "black drug" was irrelevant to finding that the 100-to-1 sentencing disparity between crack and powdered cocaine was motivated by discriminatory intent.[84] The same statistical facts that must be recognized to defuse the allegation of discriminatory effect—say, a proffered correlation between group membership and the facial target of the enforcement policy—are ignored when interpreting the meaning of acting on the facial target of the enforcement policy.

Having shown that many courts employ counterfactual-type reasoning to evaluate evidence of racial and ethnic discrimination, the following Part formalizes the counterfactual causal model in order to clarify precisely what it means to talk about race as a cause in the counterfactual sense, and what it means to define racial discrimination as the causal effect of race- or ethnicity-qua-treatment. Most federal judges are not familiar with the counterfactual causal framework presented in the next Part (although many experts employed to provide statistical services are), and I certainly do not contend that they are always consciously appealing to the formal model when they approach discrimination cases. I do contend that if confronted with the formalization, many of these legal actors would agree that it represents the type of exercise in which they are engaged when detecting discrimination.

---

[83] Litigants claiming discrimination must show that a decision-maker "selected or reaffirmed a particular course of action at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group." Pers. Admin. of Mass. v. Feeney, 442 U.S. 256, 279 (1979); *see also* Hunter v. Underwood, 471 U.S. 222, 227–28 (1985) ("Presented with a neutral state law that produces disproportionate effects along racial lines, the Court of Appeals was correct in applying the approach of *Arlington Heights* to determine whether the law violates the Equal Protection Clause of the Fourteenth Amendment: '[O]fficial action will not be held unconstitutional solely because it results in a racially disproportionate impact. . . . Proof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause.'" (quoting Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 264–65 (1977))).

[84] *See, e.g.*, McCleskey v. Kemp, 481 U.S. 279, 293 (1987); United States v. Clary, 34 F.3d 709, 713 (8th Cir. 1994) (rejecting the idea that "Congress' failure to account for a substantial and foreseeable disparate impact" of the crack-cocaine sentencing guidelines on African-Americans could violate "the spirit and letter of equal protection" because "belief that racial animus was a motivating factor, based on disproportionate impact, is simply not enough since the Equal Protection Clause is violated 'only if that impact can be traced to a discriminatory purpose.' The chain of reasoning of the district court simply will not support a conclusion or a finding that the crack statutes were passed 'because of, not merely in spite of' the adverse effect upon an identifiable group" (internal citations omitted)); David A. Sklansky, *Cocaine, Race, and Equal Protection*, 47 STAN. L. REV. 1283, 1283 (1995).

III. THE FORMAL MODEL AND RACE AS A TREATMENT

The mechanics are a bit tedious, but it is important to explicitly detail the primitives of the counterfactual model in order to clarify (1) what definition of causality is at work, and (2) what conditions must obtain in order to make meaningful empirical claims about the operation of causality if one is using the word "cause" in the counterfactual sense. Only after rigorously specifying what cause-qua-treatment means in the counterfactual model can we understand the objections to identifying discrimination as the treatment effect of race or ethnicity.

The primitives of the theory are as follows:

> U is a population of n units consisting of $u_{i \ldots n}$.
>
> Z is a treatment variable such that Z=t if u_i is exposed to treatment conditions; Z=c if u_i is exposed to control conditions (no treatment).
>
> Y is an outcome variable of interest measured on the units in U after experiencing Z=t or Z=c.[85]

The most simplistic causal question defines just two potential causes (or levels of treatment): *t* for treatment and *c* for control (some defined non-t state). Consider Z to be a variable indicating whether U is exposed to treatment, in which case Z=t, or control, i.e., no treatment, in which case Z=c. Each unit must be potentially exposable to either c or t. The variable Y defines an outcome of interest on each unit in U. The outcome variable Y is some measure of the effect of Z on an outcome on U that is, through some measurement mechanism, assigned a value.

The heart of the potential outcomes models is that we must define not just one value of the outcome variable for each unit but two: $Y_t(u)$ indicates the value of the response variable for unit u when exposed to t; $Y_c(u)$ indicates the value of the response variable for unit u when exposed to c (not-t).[86] The treatment is some occurrence with a specific temporality and so the

---

[85] The counterfactual model of causation is sometimes also called the "Rubin Model" or the "potential outcomes" framework. There is a massive amount of literature on this, but just to name a few seminal works, see, for example, STEPHEN L. MORGAN & CHRISTOPHER WINSHIP, COUNTERFACTUALS AND CAUSAL INFERENCE: METHODS AND PRINCIPLES FOR SOCIAL RESEARCH (2007); Holland, *supra* note 13, at 959; Donald B. Rubin, *Comment: Which Ifs Have Causal Answers*, 81 J. AM. STAT. ASS'N 961, 961–62 (1986); and Jerzy Splawa-Neyman, *On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.*, 5 STAT. SCI. 465 (1990). For a very clear and accessible summary, see D. James Greiner, *Causal Inference in Civil Rights Litigation*, 122 HARV. L. REV. 533, 557–63 (2008).

[86] For simplicity's sake, I follow Holland and denote the value of the outcome variable on unit u_i for the world in which the unit experiences the treatment, Z=t, as $Y_t(u)$, and the value of the outcome variable on unit u_i for the world in which the unit experiences the control, Z=c, as $Y_c(u)$. Holland, *supra* note 13, at 947.

response variable Y must be measured at some point in time after t or c is determined.

The individual causal effect of treatment t as measured by outcome variable Y (relative to control state c) on unit u is defined as $Y_t(u) - Y_c(u)$. Note the causal effect is defined as the difference in the response variable under treatment and control conditions *for the same unit*: it is the difference between the value of Y for unit u that would obtain under treatment t, and the value of Y for unit u that would obtain under control conditions. A causal effect is defined in reference to a clear counterfactual (not-t, i.e., c), hence the name counterfactual or potential outcome.

Causes in the statistical counterfactual paradigm must be, at the risk of mixing disciplinary metaphors, what philosopher David Lewis called "non-backtracking"; that is, it must be possible to imagine a treatment on the unit at a given time (time=1) that does not entail a host of other changes to the unit prior to that moment (time<1).[87] Backtracking counterfactuals recognize that there are some changes in the current state of affairs that necessarily imply that other entangled states were also changed. As Lewis explained, "if the present were different, the past would be different too."[88] By way of example, consider an attempt to measure the causal impact of Hillary Clinton—as opposed to Donald Trump—being inaugurated on January 20th, 2017 on Washington, D.C. Metro ridership. Obviously, a number of other entangled prior events—potentially ranging from voter turnout to the timing of the FBI's announcement about finding a Clinton aide's emails on Anthony Weiner's laptop—would have to have turned out differently in the past for a Clinton inauguration to be a potential cause of a change in D.C. Metro ridership. Such a question can only be specified as a backtracking counterfactual. But treatments in the statistical counterfactual model exclude backtracking. The model requires the possibility of a treated and nontreated state for otherwise identical units that could bring about different potential outcome states in the future without the treatment entailing other changes to the unit in the past.[89]

The fact that causation is defined in terms of the counterfactual experience of one unit in two alternative conditions leads to what has been

---

[87] *See* 2 DAVID LEWIS, PHILOSOPHICAL PAPERS 33–35 (1986). As Lewis wrote, "[c]ounterfactuals are infected with vagueness," and "[d]ifferent ways of (partly) resolving the vagueness are appropriate in different contexts." *Id.* at 34.

[88] *Id.* at 33.

[89] As Lewis explained the nonbacktracking counterfactual requirement, "[I]f the present were different the past would be the same, but the same past causes would fail somehow to cause the same present effects." *Id.* at 34.

termed the "fundamental problem of causal inference."[90] We only observe either $Y_t(u)$ or $Y_c(u)$, so we can never identify the true causal effect. Bummer.

There are various methodological deliverances from this apparent epistemic bummer. For purposes of our discussion here, the logic of these methods involves trying to create or identify subgroupings of U exposed to either t or c and measuring the outcome of interest, Y, on those subgroupings. The aim of the various techniques is either to construct an assignment process whereby there would be no reason to think there is a correlation between the outcome of interest and assignment to the treatment category (known as *ignorability*), or somehow to retroactively "balance" theoretically relevant variables between the treated and control subgroups. Stated with extreme simplification, one gets around the fundamental problem of causal inference either by randomly assigning treatment, or by reconstructing two groups of treated and control units that are similarly situated with respect to all relevant variables and then estimating the average treatment effect between the treated and control groups.[91]

A. Is Race a Treatment? The Rubian Statistician's Objection

Can race be a treatment—and thus a cause—in the counterfactual framework? Surprisingly, given how common it is to talk about race as a cause in this way, most statisticians working in this tradition would say no. Or, more specifically, they would say no if the units about which we are making causal inferences are the raced units (such as individuals). Although many people have voiced this objection, I am calling it the Rubian statistician's objection after Donald Rubin, one of the early innovators of the formal counterfactual causal model.[92] Most people working in this

---

[90] Holland, *supra* note 13, at 947.

[91] Quick and dirty, it goes like this: First, define the average causal effect, T, of treatment t (relative to control state c) over many units in the population, U, as the expected value of the difference $Y_t(u) - Y_c(u)$ over all the u's in U: $T = E[Y_t] - E[Y_c]$. Second, estimate what is observable: $T^* = E[Y_t|t] - E[Y_c|c]$. Third, figure out under what conditions one might think that $T^* = T$, i.e., disentangle association from causation. The logic is to use other units to fill in the missing values of the counterfactual outcomes of the units for the treatment or control each did not receive. There are a number of other assumptions necessary to draw causal inferences not discussed at length here (for example, the stable unit treatment value assumption (SUTVA), which means the treatment status of any unit does not affect the potential outcomes of the other units (noninterference) and the treatments for all units are comparable (no variation in treatment or, said another way, the outcome value for unit u when exposed to treatment t (Y_t) is the same no matter what mechanism is used to assign the unit to the treatment)).

[92] As Rubin put it,

> [W]ithin our model, each of the *T* treatments must consist of a series of actions that could be applied to each experimental unit. This requirement may seem obvious, but some colloquial uses of "cause" specify treatments that either cannot be applied or are so ambiguous that no series of

framework would say a question like "Did Eddie Murphy's race cause him to get a free newspaper?" does not have meaning in the counterfactual causal sense.

The first issue the Rubian statistician points to is that, as described above, causes are interventions with a discrete temporal dimension. A treatment (Z) is a manipulation of something *on* a unit (u) or something *to which* the unit can be exposed that we hypothesize would bring about a change in a measured outcome of interest (Y) about u. The unit must be susceptible of being subject to the two distinct states of treated and nontreated (Z=t and Z=c) such that it is meaningful to talk about the potential outcomes *of the same unit* in these two states, $Y_t(u)$ or $Y_c(u)$. The famous Rubin/Holland slogan—"No causation without manipulation"—captures the definitional requirement in the counterfactual model that causes are only those things that we can, at least hypothetically, bring about on the unit.[93] If Eddie Murphy's race is an "immutable characteristic" assigned at birth, then it is not meaningful to talk about his race as a cause of an outcome within the counterfactual causal paradigm because it cannot be intervened on at some later point.[94] Race is not a manipulative variable *on* the unit, but a trait *of* the unit of interest.[95]

---

actions can be inferred from the description of the treatment; such questions have no causal answer within our framework.

Donald B. Rubin, *Bayesian Inference for Causal Effects: The Role of Randomization*, 6 ANNALS STAT. 34, 39 (1978).

[93] Holland, *supra* note 13, at 959.

> An attribute cannot be a cause in an experiment, because the notion of *potential exposability* does not apply to it. The only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit. Statements of "causation" that involve attributes as "causes" are always statements of association between the values of an attribute and a response variable across the units in a population.

Id. at 955; *see also* Tyler J. VanderWeele & Whitney R. Robinson, *On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables*, 25 EPIDEMIOLOGY 473, 473 (2014) ("Part of the challenge of interpreting race coefficients causally is that, in the formal causal inference literature, effects are often defined in terms of counterfactual or potential outcomes, which are in turn defined as the outcomes that would result under hypothetical interventions. There are, however, no reasonable hypothetical interventions on race when race itself is the exposure." (footnote omitted)).

[94] There are many critiques of the Rubian causal framework and different approaches to causality, but I focus on this framework because it best captures the hegemonic approach in courts and social science of defining discrimination as a causal relation between a protected attribute and an outcome.

[95] The statistician's objection to thinking about race as a cause-qua-treatment is often explained (and debated) in terms of nonmanipulability: the physical impossibility of bringing about a change in the race of an individual at some point after birth. *E.g.*, JAMES WOODWARD, MAKING THINGS HAPPEN: A THEORY OF CAUSAL EXPLANATION 132 (2003); Alexandre Marcellesi, *Is Race a Cause?*, 80 PHIL. SCI. 650, 652–53 (2013). But, as explained below, the entire terms of that debate miss the real question—What notion of RACE is presupposed by a model in which it can theoretically be a treatment, and is that notion sociologically plausible?

Second, as a consequence of this, the most important other facts about the unit happen posttreatment, and therefore, there is no way to separate out the causal effect of the treatment from mechanisms or intermediate outcomes. It is meaningless to talk about, for example, a white Eddie Murphy, because every aspect of Eddie Murphy's person from birth onwards was formed by living as a black man in the United States of America. The methodological problem this poses is that we cannot separate the fact of Eddie Murphy being black from other facts about him—class, education, humor, affect, dress, speech, etc.—in order to isolate the causal force of *race alone* on an outcome of interest. To use the language of backtracking, it is hard to imagine a treatment to his person to be white at, say, age twenty-three, without backtracking a substantial number of other facts about his life that would have been different if he had been born, raised, and perceived to be white his entire life prior to age twenty-three. With respect to observational methods that try to construct comparable subgroupings, we cannot "control for" or "balance on" all of the theoretically relevant posttreatment intermediate outcomes because, said in fancy statistical language, "controlling for a post-treatment variable messes up the estimate of total treatment effect . . . ."[96]

These are the standard statistical, definitional, and methodological objections to talking about race as a cause-qua-treatment. Much of this has been stated, restated, and debated at length in the social science and statistics literature.[97] Ultimately, I want to put a sociological spin on the Rubian statistician's objection, but for now I sum up the statistician's objections to talking about race as a treatment in two points: (1) nonmanipulability: the units are not equally potentially exposable to treatment and control; and (2) temporality of treatment: the inherent confoundedness of treatment effects and posttreatment intermediate outcomes.

B. Is Race a Treatment? The Greiner–Rubin Statistician's Solution

But statisticians don't just have objections; they also have solutions. The following Section argues that the statistician's solution is an unsatisfactory model of discrimination for the same reasons that I believe the statistician's objections outlined in the prior Section have deep sociological

---

[96] ANDREW GELMAN & JENNIFER HILL, DATA ANALYSIS USING REGRESSION AND MULTILEVEL/HIERARCHICAL MODELS 188 (2007).

[97] *See generally* MORGAN & WINSHIP, *supra* note 85, at 439; D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 REV. ECON. & STAT. 775 (2011); Holland, *supra* note 13; Marcellesi, *supra* note 95, at 655; Maya Sen & Omar Wasow, *Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics*, 19 ANN. REV. POL. SCI. 499 (2016).

meaning, deeper meaning than I think they have been given in the methodological debate. But first, the proffered solution.

The statistician's solution to the objections to talking about race as a cause-qua-treatment within the counterfactual causal framework is to switch the unit of analysis from the raced unit (person, aggregation of persons) to the decision-maker confronting the raced units. James Greiner and Donald Rubin present the most eloquently formalized version of this solution, although it is the same logic informing audit or correspondence studies (insofar as they claim to identify a treatment effect of race, which is in no way a logical entailment of the method).[98] Greiner and Rubin describe the solution as a "shift in emphasis to perceptions of immutable characteristics," arguing that this "allows some well-defined causal questions to be posed and, within the limits of observational studies, inferences to be drawn," or alternatively, at least to "identify a set of assumptions that allows causal analysis."[99]

The basics of the approach go something like this: The unit is a decision-maker with power to determine the outcome of interest (Y). The decision-maker perceives a distinct candidate unit (candidates for the outcome of interest). The treatment is the "immutable characteristic" of the candidate *as perceived* by the decision-maker.[100] Thus, the unit about which we are making causal inferences is the decision-maker, which could be an individual or aggregate decision-maker such as a firm or police department. The treatment is the perceived race of a candidate unit, and the outcome of interest is some outcome (Y), over which the decision-maker has the power to decide.

The statistician's solution addresses both the manipulability and temporality of treatment objections summarized in the prior Section. It addresses the manipulability issue by varying the race of the candidate units

---

[98] For clear explanations of the decision-maker causal inference framework, see Greiner & Rubin, *supra* note 97, at 776, and MORGAN & WINSHIP, *supra* note 85, at 440–41.

[99] Greiner & Rubin, *supra* note 97, at 776. Many people embrace this as the right conceptualization for thinking about discrimination as the causal effect of race, ethnicity, or sex. *See, e.g.*, JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MOSTLY HARMLESS ECONOMETRICS: AN EMPIRICIST'S COMPANION 5 (2009); Sonja B. Starr, *Explaining Race Gaps in Policing: Normative and Empirical Challenges* 32 (U. Mich. Law & Econ. Research Paper Series, Paper No. 15-003, Jan. 2015), http://papers.ssrn.com/abstract=2550032 [https://perma.cc/7XZS-FGQA].

[100] Greiner and Rubin use the language of "perceived race," *supra* note 97, at 775, but as I will argue in the following Section on audit and correspondence studies, it is better described as "signaled race" because the treatment is using one of various ways in which the social category of race is triggered visually, aurally, or with written signs. *See* Sen & Wasow, *supra* note 97, at 509 ("[T]he best way to think about the treatment in exposure [design] studies is not as perception but instead as a signal about race. After all, in an experimental context, the researcher can manipulate the signal to which the subject is exposed but not what the subject actually perceives. Second, perceived race is rarely observed . . . .").

a decision-maker perceives as opposed to manipulating the race of the candidates themselves. It addresses the temporality of treatment issue by locating the treatment at one salient moment in time—at the time the decision-maker is appraised of the raced status of the candidate or racial composition in the case of aggregate units. Defining the treatment as the moment of racial perception allows one to control for or balance many variables theoretically related both to race and to the outcome of interest, because they can now be thought of as pretreatment.

As in any "clinching" deductive method, very strict assumptions must be fulfilled to "clinch" the conclusion of causality.[101] One assumption necessary to draw causal inferences about the treatment effect of a perceived candidate's race on a decision-maker's outcome is that there be some clear concept of what the treatment and the counterfactual to the treatment *is*. The decision-maker must have in her or his mind a unitary and discrete concept of BLACK and WHITE (if we are talking about racial binaries) that is triggered by a stimulus, rather than a multidimensional conception of RACEDNESS.[102] In addition, the treatment for all units must be identical, meaning the singular discrete concept of BLACK and WHITE triggered by the stimulus is the same across potential decision-makers.[103] The method of treatment must not affect the potential outcome, meaning how the decision-maker comes to perceive the raced status of the unit does not affect the outcome of interest.[104] Finally, the noninterference assumption requires that the potential outcome of a particular decision-maker does not depend on the treatment of candidate units assigned to other decision-makers.[105]

---

[101] Nancy Cartwright, *Are RCTs the Gold Standard?*, 2 BIOSOCIETIES 11, 12 (2007) ("Clinchers are deductive: *if* they are correctly applied *and* their assumptions are met, then *if* our evidence claims are true, so too will be our conclusions—a huge benefit.").

[102] That can be relaxed, as Greiner and Rubin point out, *supra* note 97, at 778, but I think that just pushes the envelope to depicting RACEDNESS as gradations of the solid thing RACE.

[103] There are various ways to state the requirement that, however one defines the "treatment," it must be the *same* thing to every unit to which it is administered. Some versions of this requirement include the following: there is only one version of each treatment; counterfactual states must be well-defined; the treatment of different units is comparable; and there is a "single essential, counterfactual state" of treated and control. *See id.*

[104] The value of $Y_t(u)$ (the outcome for unit u when exposed to treatment t) is the same *no matter what mechanism is used to assign the unit to the treatment.* Imagine some decision-makers interact with the units for long periods of time and others just interact for a short period of time; "[t]he critical assumption here is that how the perception is created does not matter, that is, the counterfactual potential outcome is 'stable,' invariant to the nature of the evidence on which the decider's perception is based." *Id.*

[105] "[P]otential outcomes for any unit of an experiment are independent of the treatment assignment of any other unit or population member under study." Robert J. Sampson, *Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology*, 26 J. QUANT. CRIMINOLOGY 489, 492 (2010).

The statistician's solution also means that the decision about *when* the treatment is conceptualized to occur is enormously important for deciding what variables are posttreatment and thus should not be controlled for or balanced on in the analysis.[106] For example, consider Judge Schroeder's objections to the Government's expert Dr. John MacDonald in the *Johnson* case, who analyzed 20,059 traffic stop forms filled out by deputies in the Alamance County Sherriff's Office. Dr. MacDonald attempted to estimate the "effect of being Latino" on post-stop outcomes, such as warnings, citations, or arrests, by estimating a series of logistic regressions that included all available control variables, which were limited in the administrative data.[107] Judge Schroeder found this evidence unsatisfactory because he believed the analysis did not compare sufficiently similarly situated Latino and non-Latino drivers, because Dr. MacDonald did not include a host of conceivable control variables—such as personal affect in response to the police—that would only be observable post-stop.[108] Controlling for differences among drivers that officers learn after making the initial stop is problematic under the Greiner–Rubin framework because they are potentially infected by the treatment, including even perhaps recording of seemingly objective differences like prior criminal record or furtive actions because it could enter as a post hoc justification of discriminatory treatment.[109]

---

[106] This, as I argue extensively below, is my point that the process of selecting a method for detecting discrimination is not merely a methodological question with normative overtones. It *is* the process of defining the concept, which is at once descriptive about the world and evaluative.

[107] According to the court, the forms contained the following information:

> the initial reason for the traffic stop; vehicle driver information (including the driver's race and ethnicity but not name); the enforcement action taken as a result of the stop (specifically, whether an officer issued a citation, made an arrest, issued a verbal or written warning, or made no enforcement action); whether the officer performed a search during the stop; the type of search (i.e., whether the search was based on probable cause, consented to, based on a search warrant, incident to arrest, or a protective frisk); whether a passenger was searched; and whether the officer found "contraband" (e.g., illegal drugs or weapons).

United States v. Johnson, 122 F. Supp. 3d 272, 308 (M.D.N.C. 2015).

[108] "Without controlling for these obvious, nondiscriminatory reasons for post-stop outcomes, Dr. MacDonald's statistical evidence does not prove dissimilar treatment between Hispanics and *similarly situated* non-Hispanics as to stop outcome." *Id.* at 363. The court was particularly concerned with the absence of variables measuring the "severity of the conduct" and reasons for enforcement action, saying,

> Dr. MacDonald's analysis requires the court to assume two major propositions: (1) similarity in *generic stop reason* means similarity in *the severity of the conduct* resulting in the stop; and (2) the *stop reason* (which Dr. MacDonald does not purport to measure) equates causally with the reason for the *stop outcome* (which he claims to measure).

Id. at 361–62.

[109] Thus, even on the terms of the model, there is a substantial number of important police and prosecutorial decisions (such as patrol allocation by neighborhood, or subjective assessment of severity

The point of the Greiner–Rubin solution is to circumscribe the encounter of interest to a discrete set of events at and after the decision-maker's encounter with a candidate unit and ask how the perceived raced status of a unit affects the decision-maker in that discrete encounter. The substantive upshot of the statistician's solution is that it operationalizes a definition of discrimination that necessarily excludes any historical effects of how race has structured units to be systematically different on relevant variables prior to the encounter.[110] The statistician's solution addresses the manipulability and temporality objections to conceptualizing race as a treatment. The solution also vindicates the intuition that detecting discrimination in a particular arena is not about seeking to hold a discrete decision-maker liable for the entire accumulated disadvantage between different groups defined by race or ethnicity. Rather, it is about holding a decision-maker liable for how race affected the outcome of interest in a particular discrete encounter over which she has control.

Solved!

C. Is Race a Treatment? The Sociologist's Objection

Not so fast.

Let me point out something about the statistician's solution to his eponymous objection. In the counterfactual framework, the sentence "A causes B" means "the effect of A is B."[111] The framework definitionally restricts the class of eligible causes to "things that could, in principle, be treatments in experiments," because the framework is erected to give precise operational meaning to causal statements by measuring the effects of known causes, not the causes of observed effects.[112] As I have mentioned, the Rubian objection is often presented in terms of nonmanipulability: the fact that units are not potentially exposable to being raced differently at the time of the relevant encounter. But physical, logistical, ethical, or practical human limitations to bringing about a treatment are not fatal flaws in considering

---

of conduct) that are not susceptible to a study design that could exclude post-racial/ethnic perception variables, because those variables are precisely the ones that would make the units similarly situated.

[110] Greiner and Rubin state, "Much here depends on a willingness to exonerate the decider from responsibility for prior events." Greiner & Rubin, *supra* note 97, at 777. I do see how even the paradigmatic form of invidious, or what economists call "taste-based," discrimination do not also hold decision-makers responsible for prior events, namely development of invidious tastes that were learned and inherited from a history and culture that has constructed particular groups as disfavored and excluded.

[111] Paul W. Holland, *Statistics and Causal Inference: Rejoinder*, 81 J. AM. STAT. ASS'N 968, 968–70 (1986) (emphasis omitted).

[112] Holland, *supra* note 13, at 954.

something a cause in the counterfactual model.[113] As long as it is logically and conceptually possible to talk about administering a process on a unit that brings about the proffered treated state—while still retaining the unit as the same unit—then the proffered treatment can be labeled a cause in the counterfactual framework. The Greiner–Rubin solution is trying to figure out a way that race can be analyzed as a cause within the counterfactual framework, so they switch the unit of causal inference to the decision-maker to address the manipulability and temporality of treatment issues accordingly and identify perceived race as the treatment. Here is where I want to introduce the sociological spin on the Rubian statistician's objection to talking about race as a cause in the counterfactual framework. The problem with talking about race as a cause-qua-treatment is not a problem with practical manipulability; it is a problem with the sociological and normative meaning of the proffered manipulation.

Race is a fundamental structuring institution of life chances in the United States. Inhabiting a particular racial category not only shapes the opportunities, advantages, and resources that will be available in a person's life course, it means living with a particular cultural meaning attached to one's body. It is not meaningful to talk about an otherwise identical person suddenly swapping racial status at the time of a given encounter because the raced status a person has inhabited since birth has shaped so many aspects of the person relevant to the encounter that it is impossible to disentangle those factors from the person's raced status.

There is no nonbacktracking way to specify the hypothetical "treatment of race" on a person (or aggregation) at a given moment to measure its effect on an outcome of interest (much like there is no nonbacktracking way to specify the effect of a Clinton, as opposed to Trump, inauguration on D.C. Metro ridership). If so many aspects of life are structured by ascribed racial status—from prenatal medical care, to residential patterns, to educational opportunities, to end-of-life palliative treatment—then it is nonsensical to ask of a person (or aggregate units like neighborhoods) to change his/her/their raced status, but otherwise be the exact same person (or

[113] As Heckman pointed out, we should not confuse practical problems for theoretical problems: "Holland's 1986 claim that the causal effects of race or gender are meaningless conflates an empirical problem"—identifying parameters (causal or otherwise) from hypothetical population data—"with a problem of theory"—defining the set of hypotheticals or counterfactuals; "[t]he scientific approach sharply distinguishes these two issues. One can in theory define the effect even if one cannot identify it from population or sample data." James J. Heckman, *The Scientific Model of Causality*, 35 SOC. METHODOLOGY 1, 31–32 (2005); *see also* Clark Glymour, *Comment: Statistics and Metaphysics*, 81 J. AM. STAT. ASS'N 964, 964–66 (1986); Clark Glymour & Madelyn R. Glymour, *Commentary: Race and Sex Are Causes*, 25 EPIDEMIOLOGY 488, 489 (2014) (discussing conditions under which race and sex can be conceptualized as causes in a framework limiting the label to interventions or events on the unit).

neighborhood). It is impossible (that is, illogical, nonsensical, improbable, meaningless) to ask of the person: Be the exact same unit *except for race*, but do not change anything else about yourself because I want to see the effect of race and race alone on an outcome (Y).

But the thrust of the Greiner–Rubin solution is to posit this as a possibility in the minds of decision-makers. That is, the statistician's solution requires decision-makers' consideration of candidate units that are the exact same units *except for race*, asking that those units not change anything else about themselves except that one attribute so that we might isolate the effect of race and race alone on the decision-makers' assignment of an outcome (Y). This solution requires us to accept that there is such a thing in potential decision-makers' minds as unit u that is "treated" by being raced "black" that is conceptually an identical unit when raced "white," *except for that single trait*. That is the narrow and precise definition of cause in the counterfactual model. Something cannot be a causal predicate in the counterfactual framework if

> we cannot coherently describe what it would be like for the relevant intervention to occur at all or for which there is no conceivable basis for assessing claims about what would happen under such interventions because we have no basis for disentangling, even conceptually, the effects of changing the cause variable alone from the effects of other sorts of changes that accompany changes in the cause variable.[114]

What sort of thing is "race" under the thought experiment that imagines decision-makers could perceive identical units that differ *only* in the treatment of racial status?[115] I submit the only way to get that thought

---

[114] WOODWARD, *supra* note 95, at 132. As I will argue below, this issue is fatal for any attempt to identify discrimination with the treatment effect of race because we require sociological knowledge of what sorts of things have different cultural or social meanings by group status despite being formally identical, and we require a moral theory for what sorts of things are fair or just to vary given the prior.

[115] A recent paper by Sen and Wasow suggests that the problems of causal inference with respect to race can be fruitfully addressed by conceptualizing race as a "bundle of sticks," which would be "operationalized as a disaggregable composite variable rather than a monolithic, homogenous entity, [and] the problem of manipulability can be resolved by identifying an element of race that is [both] relevant to the research question at hand and [that] can be manipulated in at least one of two ways," either by selecting auditors "from different racial categories," or by changing "traits that are highly collinear with race and mutable [which] are often well suited to causal inference" such as name, neighborhood, or dialect. Sen & Wasow, *supra* note 97, at 506–08. They argue this "approach resolves the conflict between the potential outcomes framework of causal inference and seemingly immutable characteristics such as race, gender and sexual orientation." *Id.* at 500–01. VanderWeele and Robinson, *supra* note 93, at 477, suggest a related way of interpreting the race coefficient in a proffered analysis that would regress some measure of, say, health outcomes on socioeconomic status (SES) and race, saying that

> the coefficient for black race in the regression could be interpreted as the health inequality that would remain between blacks and whites if the family and neighborhood SES distributions . . . of

experiment off the ground is to conceptualize race as only a signifier stripped of all of its accreted meanings, as something like freckles or bunions that *is* just the bodily property. On this account, race is an individual-level physical attribute, and decision-makers have the capacity to identify the treatment of racial status by perceiving the attribute without triggering the social meanings (as opposed to mere affectual distaste) that make it a culturally salient category (in contrast to freckles, which can be identified by designating small dots on a person's face without deeper relevant cultural literacy). My claim is that such an account of race as a treatment is incompatible with the constructivist account of race because race is a system of social meanings, not an individual-level attribute.[116]

An analogy to other things that are not profitably conceptualized as treatments in the counterfactual framework might be helpful. Imagine one wants to understand how the treatment of GERMAN versus URUGUAY on the units of nation-states affects some outcome Y measurable at the level of the nation-state. What does it mean to think of types of COUNTRINESS as causes in the counterfactual sense? How does one pick out that thing that is distinctly THE-TREATMENT and separate it from the set of things that can conceptually be identified as NOT-THE-TREATMENT? How would one isolate GERMAN versus URUGUAY from its confounders, by stripping away history, institutional structure, culture, language, and so much more to get to the core of GERMANNESS or URUGUAYNESS? Such a thought experiment takes us away from the real social entities that we are trying to understand and seems to misrepresent the way that country-specificity has causal properties. One could create typologies of nation-states and formalize certain qualitative differences and similarities among them to understand why Y levels work out differently in different clusters of, say, resource endowments, political histories, or institutional configurations. But that is a very different exercise from isolating the causal effect of COUNTRINESS, which requires that there *be* such a thing as COUNTRINESS that is distinct and apart from NON-COUNTRINESS and the latter needs to be stripped away to get at the core

---

the black population were set equal to that of the white population (e.g., by setting SES for each black person to levels randomly chosen from the white SES distribution).

The bundle of sticks metaphor suggests that the causal properties of the constituted category would remain invariant to individual manipulations to its constitutive elements; but if the category has the meaning and causal properties it has because of its constitutive parts and their organization, then such an assumption would not hold. The theoretical objection to the VanderWeele and Robinson approach is similar: Why would we assume that the relationship between health outcomes and the social categories of BLACK and WHITE would be the same in a world in which family and neighborhood SES distributions were radically changed from what they are in our current world?

[116] Another way of thinking about the objection is that we need a theory of what *constitutes* THE-TREATMENT in order to separate it from things that are NOT-THE-TREATMENT.

treatment so we can construct otherwise identical units that differ only between GERMANNESS and URUGUAYNESS, but nothing else. Certainly, one can proffer GERMANNESS or URUGUAYNESS as an explanans for the explanandum of variation in Y, but I would assert that when we do so we are similarly offering a constitutive explanation—referencing the things of which the nation-states of Germany and Uruguay consist to explicate why such a constellation make it possible for Y to vary in the way it does.

Decision-makers do not perceive neutral units as bearers of credentials or decision-relevant variables that are then painted over with a racial status any more than they perceive nation-states as neutral units of political territory painted over with flavors of countriness. If we accept that race is such a salient vector of social life that it is incoherent to conceptualize the causal effect of race by asking a person to be the exact same person *but for* race at a particular moment in time, then it is similarly incoherent to conceptualize the causal effect of race by imagining decision-makers to perceive two candidates as otherwise identical *but for* race for similar reasons. The catch in the thought experiment is not biology making race "immutable," but history, economics, and sociology making race a fundamental structuring category of thought, perception, action, and experience in the United States.

Of course, we *can* present decision-makers with candidates of varied racial statuses that look similarly situated with respect to formal credentials. We can even present decision-makers with two candidates defined by vectors of qualifications that represent an exhaustive list of every possible variable that could conceivably be rationally relevant to the decision in question, but that differ by racial signifiers.

What sort of information is generated by such an exercise and how does it help us identify discrimination? I begin, naturally, with Eddie Murphy.

IV. EDDIE MURPHY AND THE EXPERIMENTAL IDEAL

In 1984, Eddie Murphy pioneered the field of audit studies, and researchers have been trying to replicate his genius ever since.[117] After applying white face makeup, watching a lot of *Dynasty*, and reading copious Hallmark cards, Eddie Murphy emerged in New York City as Mr. White. In his new racial identity, Murphy-as-White was gifted a free newspaper in a bodega, treated to a musical drinking party on a public bus after the last black

[117] Just kidding—there were audit studies prior to 1984 (for example, the classic correspondence study: Richard D. Schwartz & Jerome H. Skolnick, *Two Studies of Legal Stigma*, 10 SOC. PROBS. 133, 134 (1962)). However, there is no question that none have matched the genius of that 1984 *Saturday Night Live* skit "White Like Me." NBC.COM, *Watch Saturday Night Live Highlight: White Like Me*, http://www.nbc.com/saturday-night-live/video/white-like-me/n9308 [https://perma.cc/VTF7-WBQD].

man disembarked, and—with no collateral, no credit, and no ID—effortlessly procured a loan of $50,000 from a bank with assurances from the white banker that he did not really need to pay it back.

The classic audit study design mixes the randomization of clinical experimental design with real-world conditions by staging randomized encounters to generate the outcome of interest from actual decision-makers in the field.[118] One often hears the claim that randomized experiments are the "gold standard" for causal inference because randomization eliminates selection-into-treatment bias. The strength of randomized studies is internal validity, which refers to the confidence one can have in the estimate of the treatment effect within the sample studied. But their weakness is external validity, which refers to the confidence one can have that the treatment effect estimated from the study sample can be extrapolated to the general population of interest. It requires strong assumptions to infer that the treatment effect estimated from a controlled randomized study generalizes to the larger population.[119] In audit studies, the researcher tries to select auditors that are as closely matched as possible on relevant dimensions and differ only by race (or ethnicity, sex, etc.), who are then trained to go into a particular encounter in an identical fashion to solicit the decision-makers' reactions to specific prompts. Correspondence studies mimic this design without live testers by using identical résumés or applications to solicit the decision-makers' outcome and signaling the category of difference (race, ethnicity, sex) either explicitly or implicitly.

For example, Devah Pager's well-known study of race and criminal record discrimination—first conducted in Milwaukee and replicated in New York City—was designed to present employers with similarly qualified job applicants who differed by race and felony drug conviction. The Milwaukee study enrolled college students as auditors, who were "matched on the basis of age, race, physical appearance, and general style of self-presentation" and were also assigned "fictitious résumés that reflected equivalent levels of education and work experience."[120] In Ian Ayres and Peter Siegelman's

---

[118] "While retaining the key experimental features of matching and random assignment important for inferences of causality, this approach relies on real contexts (e.g., actual employment searches, real estate markets, consumer transactions) for its staged measurement techniques." DEVAH PAGER, MARKED: RACE, CRIME, AND FINDING WORK IN AN ERA OF MASS INCARCERATION 48–49 (2007).

[119] *See* CHARLES F. MANSKI, IDENTIFICATION FOR PREDICTION AND DECISION 226–27 (2007). The tension among the extent of assumptions, the confidence in the deduction, and the scope of applicability of that deduction has been widely discussed. One notable formulation is Manski's Law of Decreasing Credibility: "The credibility of inference decreases with the strength of the assumptions maintained." *Id.* at 3. As he argues, "[t]his principle implies that empirical researchers face a dilemma as they decide what assumptions to maintain. Stronger assumptions yield inferences that may be tighter but less credible. Methodological research cannot resolve this dilemma but can clarify its nature." *Id.*

[120] *See* PAGER, *supra* note 118, at 59.

classic study of discrimination in bargaining for new cars, the auditors were chosen within a narrow age range (twenty-eight to thirty-two years of age), education (three to four years of postsecondary education), attractiveness ("testers were subjectively chosen to have average attractiveness"), wore similar "yuppie" clothing, and "drove to the dealership in similar rented cars."[121] The auditors were also given careful training and instructions so that they would approach the car dealers in the same way and systematically bargain in the same manner; a "script governed both the verbal and nonverbal behavior of the testers."[122] Other well-known correspondence and audit studies include in-person consultations and sending identical credentials on résumés mailed to job openings with different racialized sounding names (white-associated names such as Emily Walsh or Greg Baker, or African-American-associated names such as Lakisha Washington or Jamal Jones).[123]

One might think that the audit study is the methodological incarnation of the statistician's solution of conceptualizing race as a cause-qua-treatment. The object of causal inference is the decision-maker; the researcher can design auditors to present a set of identical credentials; and the treatment signals the racial status of candidates. At the risk of being repetitive, audit studies are important and valuable. But they are not a clean methodological solution to messy questions about what racial discrimination is and how we ought to detect it.

The following Sections make three related conceptual points. First, audit studies do not measure the objectively bounded treatment effect of race and race alone, stripped of all other confounding traits, meanings, characteristics, or variables, on decision-makers' determinations of an outcome of interest. Second, the results of audit studies are recognizable as evidence of discrimination by virtue of constitutive explanations that ground thick ethical evaluations. It is not because the study design has isolated the treatment effect of race (or ethnicity, or sex, etc.), but because the design instantiates moral intuitions that—given what the different social categories are—members of the respective groups ought to elicit the same outcomes

---

[121] Ian Ayres & Peter Siegelman, *Race and Gender Discrimination in Bargaining for a New Car*, 85 AM. ECON. REV. 304, 305–06 (1995).

[122] *Id.* at 306.

[123] *See, e.g.*, Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991 (2004); Judith D. Feins & Rachel G. Bratt, *Barred in Boston: Racial Discrimination in Housing*, 49 J. AM. PLANNING ASS'N 344 (1983); Stephen L. Ross, *Appendix A: Paired Testing and the 2000 Housing Discrimination Survey*, *in* ANGELA WILLIAMS FOSTER ET AL., MEASURING HOUSING DISCRIMINATION IN A NATIONAL STUDY: REPORT OF A WORKSHOP 49, 49–66 (2002), https://www.nap.edu/read/10311/chapter/9 [https://perma.cc/ZRA4-LN4U]; John Yinger, *Evidence on Discrimination in Consumer Markets*, 12 J. ECON. PERSP. 23, 33 (1998); John Yinger, *Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act*, 76 AM. ECON. REV. 881, 883 (1986).

when they present in the manner in which the study manufactures. The results are often recognized as such compelling evidence because study design frequently instantiates widely held—yet often quite minimal—moral intuitions about how groups ought to be treated given what differences and meanings constitute their status. Third, semi-experimental (or experimental for that matter) methods are not the "gold standard" for detecting discrimination against which all other means should be measured. In the context of certain forms of discrimination, these methods might often be inappropriate.

A. Solid State Race

Audit studies do not measure the objective-isolated treatment effect of race and race alone. This is not because researchers have failed to design and execute the methodology with rigor or precision. Audit studies do not measure the objective-isolated treatment effect of race and race alone because there is no such thing to measure.

If one accepts the constructivist account of race, then the signifiers used to signal treatment into a racial category for auditors inflect the very meanings and substantive relevance of other decision-relevant credentials because that's what race *is*: it is a system of social meaning wherein particular cultural cues indicate the stratifying social types operative in a particular place and time. The constructivist would insist that the treatment presented in audit and correspondence studies is not substantively identical candidates that differ only by racial status, but rather differently raced candidates bearing whatever set of formally similar credentials the researcher gives them. Before further explaining what I take audit studies to be detecting, I want to make a few points about what audit studies represent by thinking through their actual design, execution, and interpretation.

First, there is no way to even design a study attempting to present decision-makers with similarly situated candidates that differ by race without thick cultural knowledge about the social meaning of traits and credentials in a particular place and time, or without making substantive value judgments about the fair grounds for decision-making in a particular arena *given* our knowledge of how these categories are constituted. Second, presenting decision-makers with candidates signaling different racial status and an identical set of credentials does not mean that decision-makers perceive identical units but for the treatment if the very meanings of the credentials are inflected by the racial status. These credentials do not necessarily *mean* the same thing for purposes of interpreting an action as discriminatory just because they are identically proffered. That is, if the social categories of race and ethnicity are constituted by systematic

differentiation over some set of important social, economic, and cultural factors, then signaling different membership changes the meaning of other decision-relevant traits or credentials.

Most audit studies proceed by trying to make all factors about the auditors that might be theoretically relevant to the decision-maker's decision identical. Imagine that the ideal experiment would have an exhaustive list of all credentials that would be instrumentally rational to consider in the decision of interest. The researcher could then set those credentials to be the same between the two candidate auditors selected to signal different raced statuses. But the researcher cannot just use credentials to signal the relevant variables if what she is trying to do is make candidates substantively identical because formally identical credentials are interpreted differently with differently raced candidates precisely because of the social fact of race.

For instance, say a researcher wants to signal equal educational attainment of a black and white auditor. She could list a high school diploma on the résumé for each auditor. But the history of racial segregation and state neglect of black schools has created large differences in the mean quality of schools between white and black neighborhoods in most major cities, so a decision-maker might treat the high school diploma signal differently for a black and white auditor based on this knowledge or presumption. Or is the right way to set up a study design to have a black student with a high school diploma from a well-known and predominantly white high school? Would that design trigger the types of assumptions Justice Clarence Thomas famously frets about with regard to affirmative action?[124] Or is the race-neutral way of proffering the credential for the high school diploma to be from a racially integrated high school? What does that mean if only a tiny percentage of the city's schools are racially integrated or if education is stratified within schools? Or should the researcher include the score of some so-called skill or aptitude test? Is that a race-neutral way of signaling capacity when the tests measure developed abilities, which systematically differ by environments, the very environments that systematically differ by

[124] Dissenting from the majority opinion upholding University of Michigan Law School's consideration of race in admissions, Justice Thomas argued:

> As admission prospects approach certainty, there is no incentive for the black applicant to continue to prepare for the LSAT once he is reasonably assured of achieving the requisite score. It is far from certain that the LSAT test-taker's behavior is responsive to the Law School's admissions policies. Nevertheless, the possibility remains that this racial discrimination will help fulfill the bigot's prophecy about black underperformance—just as it confirms the conspiracy theorist's belief that "institutional racism" is at fault for every racial disparity in our society.

Grutter v. Bollinger, 539 U.S. 306, 377 (2003) (Thomas, J., dissenting).

race?[125] Even if the researcher arranges for all relevant observable factors to look nominally equal, testers can vary on "unobservables," those soft skills or interactional tendencies that make people just *seem* different.[126] So, in order to really accomplish substantive equality in all theoretically decision-relevant variables between candidate units, the auditor must know (1) the way in which the history and social practice of race structures access to qualifications and credentials, and (2) the content of cultural stereotypes by which racial status fills in meanings of various proxies of qualifications or instrumentally rational variables.

The point is not that a researcher cannot make reasoned judgment calls and defend them, though those defenses can only be advanced in sociological and normative terms. The point is that there is no race-*neutral* way of proffering credentials because we live in a racially stratified society. When one tries to think about how to construct identical candidates that differ only by race, one realizes it is a lot of work to strip away all of the NOT-THE-TREATMENT in order to get at the thing that is THE-TREATMENT because race is such a fundamental structuring category of social life, thought, and experience in America.

If one follows the logic of constructing substantively *identical* auditors but for race all the way down, the exercise becomes one of experimentally unmaking the social consequences of a racial order in a particular encounter. But the map to unmake the consequences of entrenched racial systems can only come from thick sociological and historical knowledge about how race patterns life chances and creates meanings. It can only come from situated social understanding about the complex system of meanings and practices that constitutes the very categories. The ideal experiment to detect discrimination in the counterfactual causal model is one in which the researcher uses this map to select credentials that zero out the average differences in relevant variables that were produced by the real lived institutions of racial orders, and that signal to the decision-maker that the assumptions or valuations she might assign to formal credentials drawing on

---

[125] *See* Christopher Jencks, *Racial Bias in Testing*, *in* THE BLACK-WHITE TEST SCORE GAP 55, 55–85 (Christopher Jencks & Meredith Phillips eds., 1998) (discussing the various forms of bias that are implicated by relying on so-called aptitude tests, including labeling and content bias, which misattributes the measurement object to purely innate intelligence instead of environmental factors supporting development or uses measurement techniques that systematically favor one group, or prediction bias, which refers to the fact that similar scores do not predict valued outcomes (i.e., grades, job performance) similarly between groups because of other systematic differences).

[126] James Heckman, among others, has long criticized audit studies for failing to consider the role of different variances between groups on characteristics unobservable to the researcher but observable to the potential decision-maker. *See* James J. Heckman, *Detecting Discrimination*, 12 J. ECON. PERSP. 101, 108–11 (1998).

cultural stereotypes are inapplicable to the given case. Remember: the definition of a treatment in the counterfactual model requires an isolated nonbacktracking manipulation on a unit that can be applied without transforming the unit into a different unit. Unless the researcher is presenting identical candidates but for the treatment in a given encounter, she is not measuring the treatment effect of race alone on otherwise identical units. She is doing something else.

And what have we gotten at after we have followed that map all the way down in this ideal thought experiment of constructing substantively identical auditors but for racial difference? My sense is the standard assumption is that at base what you hit is the brute fact of racial difference, which must mean the signifiers of racial status—e.g., "just" skin color, phenotype, or some other physical difference between bodies—stripped of all of the effects they produced in the social world prior to that encounter and bundled with a note to the decision-maker that any of the negative (or positive) signified meanings that have been attached to those signifiers are not true in this instance. And my sense is that many people have understood the legal refrain of "invidious discrimination" (or "taste-based discrimination") to mean precisely this—someone acting out of affectual distaste for the brute signifiers of race.

But conceptualizing *this* as the ideal thought experiment for detecting discrimination—looking for indications that a decision-maker has acted on mere affectual distaste for the brute signifiers of race—unmoors the exercise from both a sociologically coherent theory of race and an ethically sound theory for why the coercive powers of the state should be dedicated to detecting it and rooting it out. There is no public reason—worthy of constitutional amendments, federal and state legislation, various administrative agencies, and extensive public resources—to be concerned with individual-level dislike of certain physical attributes unless we have a theory about how social processes have constructed *those* signifiers as systemically disfavored through persistent material and symbolic processes *and* a theory about why that is wrong.

B. How Audit Studies Demonstrate Discrimination

In thinking through what audit studies are trying to get at, we need to remember that there is a difference between discrimination and irrationality or idiosyncratic tastes. Saying that a decision-maker undertook a course of action because a candidate was perceived to be white *means* something different than saying a decision-maker undertook a course of action because a candidate was perceived to have freckles. The former references a social category as the reason for action; the latter references speckles of melanin

on the epidermis. When we say "Eddie Murphy got a free newspaper because he was white," "white" is essentially a metonym for a constellation of social meanings produced through a complex history of slavery, immigration, and countless other forces, not a melanin deficit (or white face makeup in his case).[127] The signifiers of racial status cannot appear (to cultural insiders) as random aspects of physicality like freckles, about which people might have thin personal preferences or affectual responses. If signifiers of racial status *could* appear to cultural insiders as random aspects of physicality, then it would be quite mysterious why those markers of physicality predict other valuable social and economic resources.[128] And there certainly would be no moral grounds, distinct from a commitment to functional rationality or efficiency, to prohibit public and private actors from acting on their preferences for certain signifiers of racial status.

For these reasons, I insist that racial discrimination is a thick ethical concept that rests on an account of the system of social meanings that constitute race *and* a normative theory for why (and when) decisions that are based on those social meanings are worthy of moral concern. There is no reason above and beyond opposition to idiosyncratic aversions, irrationality, random meanness, or a general opposition to the structure of disadvantages (as opposed to whom and how they are allocated) to care about a decision-maker denying an opportunity or imposing a cost on a freckled candidate.[129] But there is a reason above and beyond opposition to idiosyncratic aversions, irrationality, or a general opposition to the structure of disadvantages to care about a decision-maker denying an opportunity or imposing a cost on a black candidate.[130] I believe that the sociological facts that explain *why* race is

---

[127] "Because the system of slavery was contingent on and conflated with racial identity, it became crucial to be 'white,' to be identified as white, to have the property of being white. Whiteness was the characteristic, the attribute, the property of free human beings." Cheryl I. Harris, *Whiteness as Property*, 106 HARV. L. REV. 1707, 1721 (1993).

[128] Without getting dragged into another large debate, I will say that the same critiques offered here can be applied to "taste-based" discrimination, which defines discrimination as essentially a random aversion or irrational prejudice that a person is willing to expend resources to indulge: "[I]f someone has a 'taste for discrimination,' he must act as *if* he were willing to forfeit income in order to avoid certain transactions . . . ." BECKER, *supra* note 28, at 16. A police officer may dislike people with freckles. And he might even be motivated to make an otherwise marginal arrest because he hates people with freckles so much. But having "preferences" for nonfreckled people is a fundamentally different sort of disposition than a "preference" for nonblack people. A police officer can have thin preferences about freckles (just hating the way they look), but not about the signifiers of blackness because those signifiers signify a much deeper and wider set of social meanings.

[129] *See* Post, *supra* note 30, at 15, 20 (discussing the trope of colorblindness in antidiscrimination law as being an admonishment to be "instrumentally rational").

[130] I understand critiques of racialized liberalism to be making a similar point: that we cannot define distributive justice as such without engaging how liberalism is itself premised on a racial order. *See* MILLS, *supra* note 27, at 208–09.

different from freckles are necessary to make sense of *why* the moral concern with the signifiers of racial status being a source of disadvantage is distinct from that of freckles. Essentially, the constructivist account of racial group formation is the source for both: by identifying the historical and social processes by which racial groups are constituted and maintained as categories of stratification and domination, we identify the processes and meanings we want to transform so that they no longer operate as categories of stratification and domination.

Therefore, in my view, audit studies can provide evidence of discrimination not by virtue of identifying a treatment effect, but by virtue of providing evidence of a constitutive claim that grounds a thick ethical evaluation. Furthermore, they are often recognized as providing strong evidence of discrimination not because of clean methodological rigor, but rather because of the widely shared moral intuitions—often quite minimal—that are instantiated in most designs.

Let's return to Eddie Murphy and his perfect audit study. He goes into the bodega as Mr. White and gets the free newspaper. He goes in as black Eddie Murphy and he is made to pay for the newspaper (let us assume other things about his dress, speech, and affect are the same). The counterfactual is true: Eddie Murphy *not* in whiteface and otherwise the same had to pay for the newspaper. But that counterfactual is not what identifies the action as discrimination.[131] What identifies the bodega worker's action as discrimination is that we reason about its *distinctive wrongfulness* only by referencing what constitutes BLACKNESS versus WHITENESS. We can only characterize the action as distinctly discriminatory—as opposed to random, irrational, or just an expression of idiosyncratic preferences of the bodega worker—by relying on our prior social and cultural understandings about what the categories of black and white are, what they mean, in what meanings they consist, etc. The thick moral claim—the label of DISCRIMINATION—is grounded in a constitutive explanation, answering

[131] Certainly, there are many instances where someone might grant that the counterfactual is true but think that the label of discrimination is not appropriate. Take, for example, the admissions policies at issue for applicants in *Students for Fair Admissions, Inc. v. Harvard*, where the plaintiffs' expert reasons in explicit counterfactual terms that race "caused" a differential probability of admission.

> An Asian-American applicant who was male, who was not disadvantaged, and whose characteristics result in a 25% chance of admission would have more than a 36% chance of admission if treated as a white applicant; more than a 75% chance of admission if treated as a Hispanic applicant; and more than a 95% chance of admission if treated as an African-American applicant (with all other characteristics unchanged).

Expert Report of Peter S. Arcidiacono at 7, Students for Fair Admissions, Inc. v. Harvard, 2017 WL 10442564 (D. Mass. 2018) (No. 14-cv-14176 ADB). Granting, for the purposes of argument, that there is a differential likelihood of admission among groups conditional on some set of academic measures does not answer the normative question of whether it identifies a discriminatory wrong.

questions such as "In virtue of what about the categories of BLACKNESS and WHITENESS can we understand this action to act on or rely upon that makes us want to disavow it?" or "Did the dispositional properties of this social category we know of as RACE make this act wrongful?"

The reason that most of us recognize results from audit studies showing differential treatment by decision-makers of differently raced candidates is because there is widespread agreement that differently raced persons with the specific credentials presented in the study *ought to be* treated similarly, not because proffering some discrete list of identical credentials or dress somehow makes the candidates "identical persons" except for this trait called "race."

Consider another example concerning the category of sex: if we send out a male and female auditor to apply for jobs carrying identical résumés and dressed in identical skirts, they are not identical candidates but for sex. They are differently sexed candidates (as signaled with recognizable cues for culturally presumed SEX binaries) that have the same résumé wearing the same skirt. The existence of the category SEX makes the same skirt *mean* something different for purposes of deciding if differential treatment is discriminatory; and we only know that because we have prior sociological knowledge about what SEX references in this place and time. Whether or not one interprets results showing a lower job-offer rate to the male auditor-candidate as evidence of sex discrimination in hiring turns on what one thinks is fair to expect in workplaces *given* that the category of SEX in our society currently is one in which differently sexed bodies with different (presumed) primary or secondary sex characteristics are expected to wear different types of attire. Furthermore, a position that we *ought* to recognize a lower job-offer rate to the male auditor-candidate as discrimination would be based on sociological and normative arguments advancing that, if we change what sorts of attire differently sexed bodies are expected to wear in workplaces, then we will transform the very meanings of sex in our society to be categorically less capable of producing oppression and inequality. The precondition to applying a thick ethical evaluation of discrimination to the results of an audit study is a moral position on what people are owed *given* what the category is, not some formal standard of equality that can be articulated without understanding what race or sex *is* in this time and place.

C. Gold Standards

As philosophers, criminologists, statisticians, and many others have pointed out, there is no a priori gold standard method.[132] Methods must fit questions. The selection of methods should be driven by careful and rigorous thinking about precisely what type of question we are asking, and explicit reckoning with a theoretical framework that specifies the social categories and processes at work. Several aspects of police and prosecutorial discrimination cases make audit studies an inappropriate gold standard, including the conceptual problems explained in the prior Sections, the complex organizational structures of police departments and prosecutors' offices, and the problems of external validity. Similar issues could be raised in other discrimination contexts, such as the complex organizational structures of large corporations or universities.

First, the conduct alleged to be unlawful in most police or prosecutorial discrimination cases cannot be analyzed at the individual level or even at a single level of interaction. Police departments and prosecutors' offices are large and complex organizations. Massive racial or ethnic disparities of the type alleged in *Floyd* or *Johnson*,[133] for example, are not the result of individual-level racist dispositions of beat cops or line prosecutors. Or rather, they are rarely explainable only by reference to individual-level racist dispositions of beat cops or line prosecutors. Explaining those disparities requires understanding how certain responses to problems and actions become conceivable; how observed patterns of outcomes emerge from the organizational hierarchy of the police department or prosecutor's office; how enforcement decisions are made; who issues directives; how internal rules are enforced or not enforced; how incentives of different actors at various levels of the organizational hierarchy are shaped; and how all of these factors interact.

There are some 36,000 officers within the NYPD[134] allocated between many hierarchical and vertical levels of the organization.[135] Which level would be the right unit to take as the decision-maker for purposes of

---

[132] CHARLES F. MANSKI, PUBLIC POLICY IN AN UNCERTAIN WORLD: ANALYSIS AND DECISIONS 36–38 (2013); Cartwright, *supra* note 101, at 11; Sampson, *supra* note 105, at 496.

[133] See *supra* notes 63–66, 81, and accompanying text for an earlier discussion of *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D.N.Y. 2013), and *supra* notes 1–6, 107 for an earlier discussion of *United States v. Johnson*, 122 F. Supp. 3d 272, 308 (M.D.N.C. 2015).

[134] *About NYPD*, NYPD, https://www1.nyc.gov/site/nypd/about/about-nypd/about-nypd-landing.page [https://perma.cc/8KNY-PNRH].

[135] *Bureaus*, NYPD, https://www1.nyc.gov/site/nypd/bureaus/bureaus.page [https://perma.cc/NW2J-4L73] (listing twenty-two of the organization's bureaus and explaining that "[e]ach bureau is headed by a chief or deputy commissioner who is appointed by the Police Commissioner and oversees the numerous functions of his or her divisions, units, and squads").

detecting discrimination in the counterfactual model? The plaintiffs in the *Floyd* case did not set out to show that the hundreds of thousands of stops of black and Latino residents resulted from the psychological dispositions of beat cops. Some of the plaintiffs' most compelling evidence was about the multifaceted interactions among various levels of the organization, such as how allocation decisions were made, how stop-and-frisk numbers were used as a performance metric for both beat cops and precinct commanders, and the pressures created at different levels of the organization to demonstrate productivity. The outcomes at issue in the *Floyd* case were produced by the complex interaction of allocation decisions, tactical decisions, incentives and directives at various levels of the organization, and beliefs and intentions at various levels of the hierarchy. Looking for racial discrimination as the treatment effect of race on discrete decision-making units at a single level of activity would be a deeply theoretically misguided way to determine whether outcomes that are produced in an interactive fashion through many layers of organizational directives, incentives, and discretion are discriminatory.

Police allocation decisions, for example, can never be modeled in the way suggested by the statistician's solution. If a court wants to know if race "caused"—in the counterfactual sense—more police to be allocated to a certain neighborhood, a sting operation to be sited in a particular community, or prosecutors to choose specific confidential informants (CIs) to make new conspiracy cases, it is impossible to do so utilizing the Greiner–Rubin solution. There is no time at which we could imagine the relevant decision-makers *not* perceiving the racial composition of the neighborhood, community, or the race or ethnicity of their CIs. Actors in the organizational hierarchy making tactical and allocation decisions are always doing so with full knowledge of the racial compositions of the geographic spaces to which they are allocating officers and directing specific tactics. In addition, in many jurisdictions the history of racial and ethnic segregation makes it impossible to find counterfactual units comparable along relevant vectors.[136] We are back to problems posed by the temporality of treatment issue discussed in Part II because all relevant decisions are posttreatment. Nor are perceptions

[136] Even perception of seemingly objective conditions like physical disorder are fundamentally structured by racial and ethnic understandings, meaning it is not clear what exactly it would mean to say two racially distinct neighborhoods are identical except for demographic composition. *See* Robert J. Sampson & Stephen W. Raudenbush, *Seeing Disorder: Neighborhood Stigma and the Social Construction of "Broken Windows,"* 67 SOC. PSYCH. Q. 319, 336 (2004) (showing that "social structure," namely racial composition, "proved a more powerful predictor of perceived disorder than did carefully observed disorder"). Furthermore, a city might have a few integrated neighborhoods or police precincts, but often the majority of the outcomes of concern are happening in segregated spaces. Therefore, taking integrated spaces as the benchmark for nondiscriminatory outcomes is deeply problematic because it assumes raced units will be treated the same in homogenous and heterogeneous spaces.

of crime conditions—even devoid of any explicit race descriptors—"race-neutral" input; the very conception of what sorts of social problems are assigned to the penal arm of the state is itself deeply racialized, and seemingly neutral designations transmit racial meaning.[137]

Finally, the very strengths of methods with strong internal validity are weaknesses when trying to generalize the relevance of the findings beyond the controlled parameters of the study design. Randomized clinical trials, for example—the method typically touted as the gold standard for making causal inferences of the counterfactual variety—provide strong evidence of very narrow claims.[138] The reliability of the conclusions of audit studies is similarly dependent upon the ability of the researcher to practically realize all of the requirements laid out in Part III, which in extremely simple terms demand that the candidate units have been made exactly the same but for the designated treatment. But a court evaluating an allegation of police or prosecutorial discrimination is not just concerned with internal validity, whether the treatment effect estimated is biased for the study population. A court must also be concerned with external validity, such as whether the causal claims supported by the study are generalizable to larger populations. For all of the "vanity of rigor" in randomized experiments, we find ourselves right back in the sloppy terrain of expert judgment, past experience, and reasoned qualitative discussion to decide if and to what extent the study's findings are representative of the causal structure in the larger world that is the object of our true concern.[139]

Beyond internal and external validity issues, courts also must ask if the information generated, even if generalizable, actually captures what is at issue in a claim of police discrimination. Scholars addressing the relevance of experimental data for policy debates have used the phrase "policy transfer" or "contextualization" to capture the fact that narrow causal claims

---

[137] For compelling historical documentation of how crime rates themselves were racialized from the start, see KHALIL GIBRAN MUHAMMAD, THE CONDEMNATION OF BLACKNESS: RACE, CRIME, AND THE MAKING OF MODERN URBAN AMERICA (2011), and for compelling psychological evidence of how the concept of crime triggers associations with the concept of blackness, see Jennifer L. Eberhardt et al., *Seeing Black: Race, Crime, and Visual Processing*, 87 J. PERSONALITY SOC. PSYCH. 876, 881 (2004) (showing "the extent to which Black faces are brought before the footlights of attention when the concept of crime is activated").

[138] *[I]f* all the assumptions for their correct application are met, then if evidence claims of the appropriate form are true, so too will the conclusions be true. But these methods are concomitantly narrow in scope. The assumptions necessary for their successful application will have to be extremely restrictive and they can take only a very specialized type of evidence as input and special forms of conclusion as output. That is because it takes strong premises to deduce interesting conclusions and strong premises tend not to be widely true.

Cartwright, *supra* note 101, at 12.

[139] *Id.* at 18.

are not directly translatable to larger policy questions.[140] This is true not only because causal effects might differ by subpopulations, but also because treatments—especially scaled-up, systemwide treatments—can alter the very context in which they are applied by transforming how strategic, reflexive actors and institutions interact.[141]

In sum, we should be careful not to let the apparent rigor and cleanliness of certain methods drive the substance of what questions are asked. There is no a priori reason that every question about the social world must fit into variable-based analysis or experimental logic, and sometimes we do violence to the inquiry at hand by forcing it to conform to the structure of favored methods that might seem like the most elegant abstraction (or the highest status methods). At issue in many discrimination cases is how large, complex, multilayered, and sometimes heterogeneous organizations have operated over extended periods of time and space. Audit studies, or regression analyses trying to reconstruct similarly situated units, might provide very valuable evidence of discrimination. But they certainly cannot "clinch" the conclusion that the suspect outcomes resulting from complex interactive mechanisms are or are not discrimination.[142]

A final concern is a sociology-of-knowledge point. There is no question that quantitative evidence has an exalted position in the hierarchy of knowledge production because the methods are apparently value-free and rigorously deductive (and perhaps also because the field is dominated by men). Using and interpreting quantitative methods requires a measure of expert training, and therefore, the validity of the conclusions from the

---

[140] Sampson, *supra* note 105, at 494. Another pithy way of saying it is that "a policy is not a treatment." Robert J. Sampson, Christopher Winship & Carly Knight, *Translating Causal Claims: Principles and Strategies for Policy-Relevant Criminology*, 12 CRIMINOLOGY & PUB. POL'Y 587, 591 (2013). Specifically,

> to recommend policy requires more than considering how a treatment would be expected to work across diverse locales. When one considers policy not as a randomized trial but as a change in institutional structure, it becomes clear that theory must be brought to bear for prediction. A policy is, by definition, a change in the rules of the game. As a result, "policy translation" involves both the problem of what happens when [a treatment is administered] *and* the problem of accounting for changes in organizational, political, or wider social structure when the treatment . . . scales up into official policy.

Id.; *see also* James J. Heckman, *Econometric Causality*, 76 INT'L STAT. REV. 1, 5 (2008) (noting that a key question for social scientists is "[f]orecasting the impacts of interventions (constructing counterfactual states associated with interventions) never historically experienced to various environments, including their impacts in terms of well-being").

[141] Sampson et al., *supra* note 140, at 600 ("[C]ontext implies more than an unarticulated background or boundary against which to generalize causes and effects. To contextualize is to consider an entrenched causal web that intervenes and shapes every point of an unfolding causal process, dictating the nature of incentives, opportunities, and institutional relationships that define the policy world.").

[142] *See* Cartwright, *supra* note 101, at 12.

methods is hard to dispute without technical expertise. So, one final cautionary point is that discrimination cases should not become dominated by experts fighting in methodological terms inaccessible to other actors; with superior intelligence, computers, and fancy math in the form of equations that look like Charlie Brown cursing, experts and only experts can detect some force field of discrimination inaccessible to the uninitiated. At issue in discrimination cases is always a theory of the relevant social categories and a theory of fairness—both of which require engagement with social and ethical matters—which can become obscured behind apparently methodological discourse.

CONCLUSION: WHAT IS TO BE DONE?

Bertrand Russell famously admonished philosophy to abandon the notion of causality, alleging that it survives, "like the monarchy, only because it is erroneously supposed to do no harm."[143] Nancy Cartwright responds that causal notions are essential to differentiate between effective and ineffective strategies.[144] In the context of police and prosecutorial actions, we often observe an association between negative policing or case outcomes and the raced status of individuals or aggregations like neighborhoods. But in order to know what is to be done about these associations, we must inquire into causal relationships.

The question of detecting discrimination could be posed in terms of effective strategies—What sort of causal understanding *helps* us to know if the practices at issue in a discrimination case should be changed? Conceptualizing race as treatment does not help us distinguish between effective and ineffective strategies for dismantling discrimination because it essentially asks what the effects of a racial signifier would be if the social facts of race were not what they are today in the United States.

The ideal thought experiment that captures the treatment effect of race and race alone asks the following: produce for the decision-maker's consideration two units with identical credentials (What if the entrenched systems of racial stratification were not so?) and purge the racialized meanings that inflect different values to similar credentials or attributes (What if racial identification did not actually change the way decision-makers perceived or evaluated formally similar things about individuals or neighborhoods?). The question is, Does information about a causal link of *that* variety pick out practices that the Constitution prohibits? I say no

[143] Bertrand Russell, *On the Notion of Cause*, 13 PROC. ARISTOTELIAN SOC'Y 1, 1 (1912).

[144] Nancy Cartwright, *Causal Laws and Effective Strategies*, 13 NOÛS 419, 420 (1979).

because that thought experiment would produce information about a particular decision-maker's idiosyncratic distaste for floating signifiers.

I contend that the familiar refrain that defines discrimination as an action or practice that happens "because of race" does not identify a relation of counterfactual causality. Instead, it identifies a constitutive relation that grounds a thick moral evaluation, which means we can only identify the distinctive wrongfulness of the action or practice by reference to what social types such as BLACK, HISPANIC, or WHITE culturally reference, in what it consists to name someone such a type, and other ways of identifying what the categories *are*.

To illustrate how constitutive, and not counterfactual causal, explanations are at work in identifying discrimination, let us return to the visiting anthropologist on the island nation stratified by Royals and non-Royals. Asking about counterfactual causal dependence is just an unhelpful way of figuring out whether stepping off the street when Royals approach is properly described by a thick ethical term such as "non-Royal debasement." And it is an equally unhelpful way of identifying effective strategies, namely what needs to be changed in order to dismantle Royal-based stratification.

We can address those questions with constitutive explanations. One would need to ask how it is possible for non-Royals to feel compelled to step off the sidewalk by reference to the constitutive aspects of the socially constructed category Royal, namely by detailing the structure and content of the social meanings and relations that make the category what-it-is. One would have to ask if the fact of non-Royals deferring sidewalk access is conceptually or logically dependent on the very structure of the social kind Royal as it currently exists. Said yet another way, if a researcher *were* able to make a person on the street of this society perceive two identical walkers *but for* purple-cape-wearing and stick-carrying—and significant other meanings about this person's actions, credentials, or behavior were left unaffected by this manipulation—then there would not *be* such a thing as ROYAL in the way this society currently knows it. There would not be a morally salient issue called "non-Royal debasement" to be addressed.

The constitutive explanation grounds the thick ethical evaluation of the act. To interpret the act of stepping off the sidewalk as non-Royal debasement (or conversely, respectful Royal obedience)—as opposed to a spontaneous adjustment to scarce sidewalk space or expression of a preference for road-walking—we need access to sociological and anthropological knowledge about what constitutes the relevant social kinds in this society. And describing it as non-Royal debasement (or conversely, respectful royal obedience) is not merely disapproving or approving of the act. It is invoking a thick ethical concept, which simultaneously *describes*,

with textured, system-level information, and *evaluates* the object to which it is applied. These two facets cannot be separated because the evaluative aspect—the expression of judgment about the act—can only be activated using the descriptive component—the constellation of situated social meanings and cultural constructs referenced by the concept.[145]

We can describe something as *discrimination* only if it implicates social meanings in a way that constitutes some social kinds as degraded or disfavored, over many domains and times. Race does not have effects in the world by triggering mere affectual dislike for random physical signifiers. In fact, I contend that our culture's signifiers of racial groups are just not available to be the objects of thin preferences (or "tastes") the way that other aspects of physicality would be, such as freckles or bunions, because of the history of racial group construction. And that same process that constructed racial categories explains why so many people seem to have the same affectual response to the same signifiers—why racial discrimination is a pervasive practice in many domains in the way that freckle or bunion discrimination is not.

We can still seek to detect discrimination using audit studies, regression techniques with observational data, and many of the same methods folks have long used in social science and legal challenges. But we should be very clear what we are doing with those methods. The argument I have advanced in the preceding 29,807 words is that what we are *not* doing with those methods is detecting the treatment effect of race in the counterfactual causal sense.

Other careful thinkers committed to the counterfactual causal definition of discrimination have thoughtfully engaged questions regarding the design and interpretation of quantitative measures of discrimination, arguing that researchers ought to be both reflective and explicit about which variables are included and excluded when trying to isolate the causal effect of race.[146] But central to all counterfactual causal accounts of racial discrimination is the notion that there *is* a solid state race in units (individuals, neighborhoods,

---

[145] The valuation is fundamentally structured by and premised upon a socially and culturally conditioned set of understandings. Thus, the normative component is only accessible to those with the linguistic and cultural competencies to decode the factual descriptive component in its social context. Abend, *supra* note 21, at 148.

[146] Starr, *supra* note 99, at 32–33; Greiner & Rubin, *supra* note 97, at 775. Ian Ayres and Jonathan Borowsky criticize the inclusion of individual police officer characteristics as an illegitimate explanation for racial disparities in policing and warn of "included variable bias." IAN AYRES & JONATHAN BOROWSKY, A STUDY OF RACIALLY DISPARATE OUTCOMES IN THE LOS ANGELES POLICE DEPARTMENT 13 (2008), https://www.aclusocal.org/en/racial-profiling-lapd-study-racially-disparate-outcomes-los-angeles-police-department [https://perma.cc/R5XU-2EMQ] (report prepared for ACLU of Southern California).

etc.), an objective fact about the units that can be isolated after stripping away all confounders. For something to be a treatment, there must be a way to pick out what THE-TREATMENT is—distinct and apart from all of the things that are NOT-THE-TREATMENT so that we are sure we are talking about identical units that differ only on the-treatment. If we cannot pick apart THE-TREATMENT from NOT-THE-TREATMENT, then we are not estimating a treatment effect of race and race alone when we compare the outcomes of candidates with some list of similar credentials and signals for different racial categories. We are doing something else.

I believe that what we are doing with *both* observational and audit studies of discrimination is building a case, collecting evidence to support that case, and otherwise "vouching" for a particular constitutive claim with moral dimensions: that a specific action, practice, or policy is possible because of the social fact of race (or ethnicity) in a manner that implicates constitutive aspects of the category that we would like to change.[147] Because racial discrimination is a thick ethical concept, the way we figure out if a specific action, practice, or policy is possible because of the social fact of race is inextricably intertwined with the grounds for the moral evaluation of whether or not it ought to be tolerated.[148]

An implication of my argument is that the disparate treatment versus disparate impact binary, so central to so much of antidiscrimination law and literature, is not a tenable distinction along the lines it has often been advanced. We cannot define the former as an outcome caused by race (or where race was a substantial motivating factor, or other "close enough" formulations) and the latter as an outcome caused by a facially neutral consideration that just happens to affect racial groups unequally.

I just do not see any difference between disparate impact and disparate treatment that can be gotten at with value-free notions of counterfactual causality, much less a distinction between classification on the basis of race as such in contrast to race-neutral factors that just happen to produce dissimilar racial impacts. Disparate treatment is often distinguished from disparate impact with reference to intentional discrimination, but

---

[147] Cartwright, *supra* note 101, at 12.

> Methods [that vouch] are more wide-ranging but it cannot be proved that the conclusion is assured by the evidence, either because the method cannot be laid out in a way that lends itself to such a proof or because, by lights of the method itself, the evidence is symptomatic of the conclusion but not sufficient for it.

Id.

[148] Constitutive claims must also "support a counterfactual claim of necessity, namely that in the absence of the structures to which we are appealing the properties in question would not exist. But the kind of necessity required here is conceptual or logical, not causal or natural." Alexander Wendt, *On Constitution and Causation in International Relations*, 24 REV. INT'L STUD. 101, 105–06 (1998).

the explanandum (discrimination) in the explanans (discriminatory intent) is not saved by reference to intent or motive (to discriminate) because the point of the explanatory endeavor is to specify which sort of purposive differentiating practices are *discriminatory* and which sort are permissible.

Furthermore, it is not clear why reliance on constitutive elements of a category (black-boxing how we determine in what those consist) can be coherently referenced as "race neutral" for purposes of deciding if use of such element is discriminatory. Calling these "race neutral" is coherent if one subscribes to a biological conception of race, in which the category consists in sharing some genetic or biological facts (but, as argued extensively above, then we have another problem, which is explaining why we need to super-size scrutiny when the state classifies on the basis of those facts).[149] But if one subscribes to the constructivist theory of race, one must recognize *some* set of cultural performances, social practices, and institutions that constitute the system of social meanings of the racial or ethnic category. If the set of racially constitutive cultural performances, norms, meanings, social practices, or institutions were empty, then there would not *be* a salient category capable of producing discrimination. There certainly would be a group of people with certain physical traits (just like there is a group of people with bunions or freckles), but there would not be a complex of social meanings such that we could talk about groups being discriminated against in the thick sense. Again, someone could reject the constructivist theory of race and hold that it is a biological or genealogical fact. But such a view simply leaves the proponent no way to distinguish the thick meaning of discrimination—a morally problematic way of allocating benefits and burdens—from mere choosing based on idiosyncratic tastes or random meanness.

We often lose sight of the practices and meanings that constitute the very categories of race because one of the properties of this social category is to appear as a natural fact about bodies instead of the effect of persistent social stratification and meaning-making.[150] But the categories of WHITENESS or BLACKNESS are only available as a basis for perceiving and

[149] In what sense, for example, would one say that "pink" is a facially gender-neutral criterion, especially if we are undertaking that analysis for purposes of asking if a state early-childhood-development program only open to children that had never been dressed in pink discriminates on the basis of sex? I can only see that claim being defended from the premise that gender is a category that is constituted by genetic and biological facts.

[150] That is, if one accepts the constructivist position on race that social practices constitute bodily signifiers as salient and meaningful, absent these there would be no such social category, or it would have a different content; "the visual salience of race comes less from any obvious physical differences and more from how social practices train individuals to look differently on certain bodies." OBASOGIE, *supra* note 10, at 62.

acting upon in a discriminatory manner because of the system of social meanings and practices that bring about the very category; said another way, properties and structure do not exist independently of each other.

Many theorists and commentators have argued in distinct fashions that antidiscrimination law ought to be a project of cultural reconstruction. For example, Robert Post has argued that antidiscrimination law should not be thought of as obliterating salient differences of race or sex, but changing the meanings; with respect to sex, Post urges an interpretation that "would not require us to imagine a world of sexless individuals, but would instead challenge us to explore the precise ways in which Title VII should alter the norms by which sex is given social meaning."[151] Reva Siegel and Jack Balkin have characterized the antisubordination tradition as "[t]he moral insistence that the low be raised up—that the forces of subordination be named, accused, disestablished, and dissolved—is our story, our civil rights tradition."[152] Andrew Koppelman proposes that what he terms the "antidiscrimination project" is necessarily an endeavor in which the state actively undertakes the goal of cultural transformation that "seeks to reconstruct social reality to eliminate or marginalize the shared meanings, practices, and institutions that unjustifiably single out certain groups of citizens for stigma and disadvantage."[153]

I concur with the content of those accounts in terms of a prescriptive vision for antidiscrimination law. However, the conceptual points that I have argued above have two important implications for antisubordination theory. First, in order to defend a position that the Equal Protection Clause *ought* to be understood as remedying group inequality, one needs a theory of what constitutes GROUPNESS in the relevant respects. Only with this in hand can we account for why we care about members of specific groups occupying disadvantaged positions in the social hierarchy above and beyond caring about the existence and shape of the hierarchy. Second, if one accepts the constructivist account of racial groups, then transformation of the constitutive meanings of the relevant groups is all antidiscrimination law

---

[151] Post, *supra* note 30, at 17, 20 ("[A]ntidiscrimination law is itself a social practice, which regulates other social practices . . . . It is because the meaning of categories like race, gender, and beauty have become contested that we seek to use antidiscrimination law to reshape them in ways that reflect the purposes of the law.").

[152] Balkin & Siegel, *supra* note 47, at 17. Reva Siegel's "sociohistorical" perspective on antidiscrimination law takes account of "preservation-through-transformation," that is, how "[i]nequality in the distribution of material and dignitary goods among groups is periodically contested, and when the legitimacy of a particular distributive regime is successfully challenged, status-enforcing practices often evolve in rule structure and rationale . . . ." Reva B. Siegel, *Discrimination in the Eyes of the Law: How "Color Blindness" Discourse Disrupts and Rationalizes Social Stratification*, 88 CALIF. L. REV. 77, 83 (2000).

[153] ANDREW KOPPELMAN, ANTIDISCRIMINATION LAW AND SOCIAL EQUALITY 8 (1996).

could coherently be about. Therefore, antisubordination can make a more forceful claim as the *only* sound interpretation of antidiscrimination norms for those that reject a biological definition of race.