Law & Society Review (2024), 1–26
doi:10.1017/lsr.2024.36

ARTICLE

Special Issue on New Perspectives on Empirical Methods and Critical Race Theory

What Is Perceived When Race Is Perceived and Why It Matters for Causal Inference and Discrimination Studies

Lily Hu[1] and Issa Kohler-Hausmann[1,2]

[1]Department of Philosophy, Yale University, New Haven, CT, USA and [2]Yale Law School, Yale University, New Haven, CT, USA

Corresponding author: Lily Hu; Email: lily.hu@yale.edu

(Received 21 November 2023; accepted 23 May 2024)

Abstract

Quantifying the causal effects of race is one of the more controversial and consequential endeavors to have emerged from the causal revolution in the social sciences. The predominant view within the causal inference literature defines the effect of race as the effect of race *perception* and commonly equates this effect with "disparate treatment" racial discrimination. If these concepts are indeed equivalent, the stakes of these studies are incredibly high as they stand to establish or discredit claims of discrimination in courts, policymaking circles and public opinion. This paper interrogates the assumptions upon which this enterprise has been built. We ask: what is a perception of race, a perception of, exactly? Drawing on a rich tradition of work in critical race theory and social psychology on racial cognition, we argue that perception of race and perception of other decision-relevant features of an action situation are often co-constituted; hence, efforts to distinguish and separate these effects from each other are theoretically misguided. We conclude that empirical studies of discrimination must turn to defining what constitutes just treatment in light of the social differences that define race.

Keywords: causal inference; race; discrimination; policing; methods

Introduction

Causal inference has come to occupy an exalted position within social science generally, and more recently, within the empirical study of law (e.g., Greiner 2008; Grossman et al. 2023; Ho and Rubin 2011). Quantifying the causal effect of *race* is one of the more consequential – and controversial – instances of this causal turn, for both conceptual and political reasons. On the conceptual front, methodologists have long debated if and how race can be designated as a treatment (i.e., a cause) within the standard causal inference framework (Glymour 1986; 2014; Heckman 2005; Holland 1986). These methodological debates have high legal and political stakes. Social scientists and legal actors – including some Justices on the Supreme Court – claim that the legal concept of "disparate treatment" racial discrimination is defined as an outcome *caused* by race.[1] Many causal inference practitioners, whom we respectfully call "causal inferencers" in this article, take this definition of discrimination to mean that the relevant racial cause is a *racial perception*. Such inferencers claim they can empirically verify or discredit claims of (at least one form of) discrimination by identifying the causal effect of racial perception.[2] As such, they use terms such as "racial bias" and "racial discrimination" interchangeably with the causal effect of race perception (e.g., Starr 2016: 501; Gaebler et al. 2022: 28).[3]

Causal inferencers have built a veritable cottage industry of methods papers and empirical studies devoted to isolating the causal effect of race perception. This industry has also been influential outside of the academy. Warring causal inference experts wielding complex statistical methods feature prominently in many legal disputes of discrimination, perhaps most notably in the recent case that ended affirmative action, *Student for Fair Admissions v. President and Fellows of Harvard College* ("*SFFA*").[4] Causal inference is sure to continue to play a role in its aftermath in the upcoming battles about whether institutions are complying with the ruling – battles which will come to define what the vaguely worded Supreme Court decision prohibits in the first place (Kohler-Hausmann 2024i). A cadre of experts will most certainly also be employed in the cascade of litigation unfolding in the wake of *SFFA* in K–12 and higher education and beyond (Starr 2024). Causal inferencers who work on legally protected statuses such as race and sex are thus positioned to play a pivotal role in not only adjudicating particular cases of discrimination but in defining the scope and content of discrimination law writ large.

This paper interrogates the assumptions upon which this enterprise has been built. We want to make clear at the outset that our argument is *not* that it is impossible or misguided to set up empirical tests of discrimination. Rather, our argument is that any effort to do so must be premised upon substantive assumptions in two areas: first, a sociological theory of what race is and second, a normative theory of what is fair and just treatment in light of what race is. Causal inference studies about race perception have rarely, if ever, acknowledged these sociological and normative assumptions, let alone defended them. As a result, they present – falsely, in our view – with the force of objective science that proceeds deductively via value-free analysis of data. Because the methods themselves are complex and expressed in highly technical notation and mathematical formalisms, they are often hard to understand, much less critique, by the uninitiated. No doubt that the appearance of the arcane also contributes to their ideological power.

This paper seeks to lay bare the assumptions that undergird the causal inference methodology at work in race perception studies, assumptions which are often obscured by the technical machinery. We do so by drawing on insights from work far afield from statistics, econometrics, computer science and the quantitative social sciences broadly conceived. Technical disciplines such as these have much to say about what these methods can and cannot achieve – but only if they are brought into dialogue with other areas of inquiry dedicated to theorizing race and racial discrimination. In particular, we will draw on critical race theory (CRT) to examine the sociological and normative assumptions upon which the prevailing framework of causal inference

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

about race has been built, and perhaps more importantly, to shed light on how to construct a better paradigm for empirical studies of race perception and discrimination moving forward. We call on law and society scholars who engage in such empirical work to make explicit the assumptions on which their causal inference exercises rely. In particular, assumptions about the cognitive content that is triggered in the minds of decision-makers when they are treated with a "perception of race" are sociological in nature and, therefore, subject to empirical verification. Other assumptions about which causal contrasts illuminate "discrimination" or "disparate treatment" are normative and conceptual in nature and require argumentation in those veins. Our aim in this article is to lay bare those assumptions; the next step, which we leave for future work, is to interrogate and substantively defend them.

"Race" in causal inference

Causes as difference-makers in the potential outcomes framework

We start by briefly introducing the widely shared concept of *causation* that undergirds dominant approaches to causal inference in statistics and quantitative social sciences. We frame our discussion within the potential outcomes (PO) framework, a leading school of causal inference research.[5] The PO framework takes the concepts of *units, treatments* and *potential outcomes* as its primitives using the following notation. Units i are the constituents of the population of interest in the study. The causal effect of a treatment D is defined as the difference that obtains on a unit with respect to some outcome Y across different treatment states of the unit: for example, in the case of a binary treatment, the difference between the outcome that obtains in a world in which the unit receives treatment set at one level ($D = 0$) and another world in which the unit receives treatment set at a different level ($D = 1$). The treatment is the cause under study. Units are the entity types that receive treatment and about which we are drawing inferences. The outcomes that obtain under different treatment settings are called the "potential outcomes."

The metaphysical account of causation that underlies this framework is that of *counterfactual dependence.* Roughly, the idea is that D is a cause of Y just in case a special kind of dependence holds between these two features of the unit – informally, that if the unit were (counterfactually) treated with a different level of D, then a different level of Y would obtain. A counterfactual analysis of the causal effect of X on Y is essentially contrastive: it asks what unit i's value for Y would be had D taken the value of, for example, 1 *rather than* 0 (Rubin 1974; Schaffer 2005).

In formal PO notation, the *individual causal effect* (ICE) is defined as:

$$Y_i (D = 0) - Y_i (D = 1) \,. \tag{1}$$

Of course, we only observe what happens in our world where D takes on its actual value – say, $D = 0$ – and do not observe what happens when D has a different value – say, $D = 1$. This "missing data" from the unobserved counterfactual about unit i generates what is known by methodologists as the "fundamental problem of causal inference" (Holland 1986: 947). The promise of causal inference is that despite these "empirical constraints on our access to deep metaphysical facts" (Paul and Healy 2018: 324), under certain assumptions, one may nevertheless infer what would have happened on average in these units' unobserved (indeed, unobservable) POs by leveraging observations about *different* units' outcomes under that treatment level.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

For this reason, researchers target quantities that average over ICEs within a population of units. This aggregate quantity is called a *causal estimand*. One popular target estimand is the *average causal effect* (ACE), expressed formally as:

$$\mathrm{E}\left[Y_i\left(D=0\right)-Y_i\left(D=1\right)\right] \tag{2}$$

where E denotes the mean or expected value of Y values, taken over the population of units.

In any given instance of causal inquiry, we can neither understand the meaning of this formal expression, much less conduct causal inference to estimate it, without specifying what the variables are meant to pick out in our world. Since this paper is primarily concerned with causal inference exercises about *race*, we now turn to how the literature conceptualizes race as a cause within the PO framework.

(How) Can "race" be a cause?

There is a longstanding debate among causal inference methodologists and practitioners about the status of race as a cause. Some working in the PO framework take causal claims to be claims about the measured effect of possible interventions and thus restrict the class of eligible causes to "things that could, in principle, be treatments in experiments" (Holland 1986: 954). Meanwhile, other scholars contend that physical, logistical or ethical hurdles to carrying out a manipulation on some variable have no bearing on whether it may be identified as a cause. These scholars maintain that race causal quantities may be defined by reference to some imagined "intervention" that simply *sets* a race variable to take some value (e.g., Pearl 2010; Heckman 2005: 31–32; Glymour 1986; Glymour and Glymour 2014).

But despite these ongoing disputes about the status of race as a cause, one interpretation of the "causal effect of race" has been widely accepted as providing a well-defined estimand: the causal effect of the *perception* of race. On this view, race is not conceived as a treatment administered to a particular individual who undergoes some sort of racial transformation. Rather, the treatment is defined as a *racial perception*, which is administered to a decision-maker who, thanks to receiving some cue about a candidate for some outcome, "perceives" the candidate to be a member of a particular racial group or racialized under a particular status (Gaebler et al. 2022; Crabtree et al. 2023; Greiner & Rubin 2011). This literature uses the term "perception of race," and we will follow that convention. However, it should be clear that what they and we mean by that term is not the mere registering of some cue or stimuli, but rather the decision-maker's cognizing race or forming racial beliefs.[6]

To illustrate how inferencers conceptualize the treatment of race perception, consider a study about prosecutors' charging decisions. The race treatment is the prosecutor's perception of race: whether the individual whose file they are assessing is perceived to be racialized white ($D=w$) or Black ($D=b$).[7] Let X be the set of other, so-called "non-race" features of the situation that the researcher has designated as causally relevant to the outcome. A unit's PO is a function of the race treatment and the "non-race" features of the unit: $Y\left(D,\ X\right)$. In this case, the POs Y are No Charges Filed ($Y=0$) and Charges Filed ($Y=1$).

The PO model defines the ACE over many units of race on prosecutors' decisions as the difference in charging outcomes that obtains across two "worlds" for each unit:

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

one in which the prosecutor perceives the arrestee whose file they review as racialized white versus a world in which they perceive them as racialized black. In formal PO notation, this estimand is written as[8]:

$$\mathrm{E}\left[Y_i\left(D=b,\ X=x\right)-Y_i\left(D=w,\ X=x\right)\right]. \tag{3}$$

The PO terms above indicate that the unit i is characterized by covariates $X=x$ under both race treatments $D=b$ and $D=w$. The idea here is that everything else about the individual and case – all their so-called "non-race" features – are the same across the two cases.

In the next subsection, we discuss why despite race perception's being the dominant operationalization of race-qua-cause across a broad range of social scientific domains, there remains a critical ambiguity in precisely what the target of these endeavors is.

The "holy grail": isolating race perception

The "race effect" has been described by the economist Roland Fryer as the "holy grail" of labor economics on discrimination; it is "the parameter that we are all attempting to estimate but never quite do."[9] Many social scientists beyond economists, including political scientists, sociologists and psychologists (e.g., Pager and Quillian 2005; Pedulla 2014; Quillian et. al. 2017), have also engaged in quantitative and quasi-experimental empirical projects with a "desire to estimate the causal effect of race – or perceptions thereof – on decisions" (Grossman et al. 2023: 94). Other common locutions explain that the quantity of interest in these studies isolates the effect of race *from* the effects of other designated non-race factors (Guryan and Charles 2013: 424) or "varying race [or gender] but keeping all else constant" (Heckman 1998: 102; see also Block et al. 2021: 1; Betrand and Duflo 2017: 310).[10] Audit and correspondence studies are often touted as the "gold standard" to do so (Pager and Shepherd 2008; National Research Council 2004; Quillan et al. 2017).

Although many causal inferencers seem to be after this holy grail of race-causal estimands, the sacred object of this crusade is blurry. It is unclear what counterfactuals define it. Furthermore, unclarity about what the holy grail *is* in turn generates unclarity about its claimed normative or legal significance. These ambiguities arise because many causal inferencers state two distinct sets of commitments. First, they state a commitment to identifying a theoretical estimand that is defined by varying the treatment "perception of race" and holding constant everything that is not a part of/entailed in this treatment. Second, they state a commitment to measuring a legal normative phenomenon: "disparate treatment" discrimination. These commitments can be at odds with each other, depending on how one defines each of those concepts.

Many causal inferencers say that these two commitments are identical.[11] Furthermore, they say that they set out to detect the causal effects of race perception *because* that is what they believe (mistakenly, in our opinion) the law defines as "disparate treatment" discrimination. For these researchers, a study accurately measures "disparate treatment" discrimination just in case it identifies the unconfounded effect of race perception. Researchers must therefore make explicit their substantive empirical theory of *what is entailed in race perception* – that is, an account of what is a part of as opposed to distinct from race perception. They must also defend

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

the content of this assumption. After all, this assumption is what allows them to properly characterize something as a confounder of the sought after causal effect as opposed to part of it. Furthermore, on their reconstruction of the law, they can label this causal effect an instance of a legal-normative category only if *that* causal effect is the effect of race perception – both in its entirety and unconfounded by other things.

Despite the fact that many inferencers say that they believe their commitment to measuring the causal effect of race perception and their commitment to measuring "disparate treatment" racial discrimination are identical, we observe these two commitments pulling in different directions in some studies. Sometimes, a researcher attempts to methodologically strike out the causal effects of a feature (i.e., treat it as a confounder) even while the researcher themselves seems to take it to be a part of what is perceived when someone is successfully treated with "perception of race." But if a researcher takes a particular feature to be a *part of* the treatment perception of race, then that feature should, methodologically speaking, *vary* across differently racialized candidates. On the other hand, if a researcher takes a particular feature to be distinct from and a potential confounder of the treatment, then it should *not vary* (and should instead be made "identical" or otherwise "controlled for") across the different race perception conditions. Treating a feature that is a part of the treatment as a confounder would make for a "bad control" (Angrist and Pischke 2009; Imbens and Rubin 2015)methodologically if the inferencer were being driven exclusively by identifying the causal effect of "perception of race" (commitment one).

But treating a feature that is part of race perception as a confounder would be methodologically justified if inferencers prioritize studying discrimination (commitment two) and the full and unqualified causal effect of race perception does not define what constitutes "disparate treatment" discrimination. Said another way, some causal inferencers appear to be methodologically driven by their substantive normative views about what kinds of similarities across differently racialized candidates entitle them to equal treatment. Here, it is a normative theory about what constitutes discrimination rather than a sociological theory about what constitutes perception of race that explains why certain effects of race perception are "discriminatory" and not others. The problem is that researchers in this camp often obfuscate the role of normative theorizing in their work. They claim to be measuring all (and only) causal effects of race perception and labeling those (and only those) "discrimination." But they are, in fact, operationalizing an unstated substantive normative theory of what constitutes discrimination, such that they label effects the "effect of race perception" just in case it is the effect of a racial perception that *they think* is discriminatory if acted upon.

The following passage from an article published in the *Journal of Empirical Legal Studies* illustrates this ambiguity and thus, the stakes of clarifying it. We choose this passage not because these authors' claims are out of the ordinary, but because they so clearly illustrate the tension between the commitment to studying the causal effects of race versus discrimination and the ambiguity about what is assumed to be part of "race perception" as opposed to cofounders of it. The authors, Grossman, Nyarko and Goel, describe the typical approach of using causal inference analysis to detect disparate treatment as being "motivated by a desire to estimate the causal effect of race – or perceptions thereof – on decisions." They go on to say:

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

> Implicitly, this design embraces a narrow definition of discrimination as *disparate treatment*: the researcher wants to know, for example, whether a Black defendant is **treated differently from a white defendant *because of* their race**. For instance, the researcher may be interested in [differences in the "Black" and "white" regression parameters] as a measure of the racial gap in decisions among **similarly situated individuals**. The primary statistical concern in these studies is omitted variable bias. Hence, it is typical for studies in this setting to include as many observable controls as possible in X_i. By adjusting for a large number of factors, the hope is that the design allows for the conclusion that differences in outcomes can be **traced back to differences in the defendants' race as opposed to other dimensions, such as criminal record or socioeconomic status**. (Grossman et al. 2023: 94) (*italic* emphasis in original, **bold** emphasis added)

In this passage, the causal inferencer assumes either that (i) perception of race does *not* entail perception of anything about (e.g.) criminal record or socioeconomic status (SES) or (ii) perception of race *does* entail some kind perception regarding criminal record or SES, but such perceptions must be struck out of the causal effect of race.

Let's examine each in turn. The first interpretation (i) makes an empirical assumption that a perception of race does not entail beliefs about SES and criminal record. In our view, this assumption is implausible, and we leave to "Assumptions about race perception and their (lack of) justification" section a more extensive discussion of why. For now, we will simply note that studies have shown that exposing subjects to "Black names" triggers beliefs about both SES (Fryer and Levitt 2004; Gaddis 2017; Simonsohn 2016) and criminal records (Agan and Starr 2018; Doleac and Hansen 2016; Holzer et al. 2006). Our main point at this time is that, per their commitment to studying the effects of race perception, whatever perceptions are part of the treatment "perception of race" are thereby part of what it is to be "treated differently from a white defendant *because of* [the candidate's] race" as Grossman *et al.* put it.

The second interpretation (ii) makes the opposite empirical assumption – that race perception *does* entail some kind of perception about SES and likely criminal record – but then describes a methodological commitment to cancel out part of the causal effect of race perception. If beliefs about SES and criminal record are entailed in race perception, then why do these authors claim that these effects must be struck out of the measured effect? The reason, evidently, would be a commitment to make the differently racialized defendants "similarly situated" in some substantive way such that dissimilar treatment would thus be discriminatory. Unfortunately, they do not tell us what notion of "similarly situated" would justify treating some constituents of race perception as confounders.

Thus, here we see that these two commitments – to identify the causal effect of race perception and to measure "disparate treatment" – might be at odds, depending on how the inferencer defines what is entailed in race perception and what constitutes "disparate treatment" discrimination. When inferencers are not explicit about the answers to these questions, it is unclear what drives their methodological choices or authorizes the inference from observing a causal effect to concluding that it is an instance of the legal-normative category "disparate treatment." For example, despite claiming that an outcome's being caused by race perception is necessary

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

and sufficient for it to count as "disparate treatment," researchers often use methods to strike out or correct for certain perceptions that are, in our view and sometimes in their own view, *entailed* in race perception. In these cases, researchers' interest in operationalizing their preferred substantively normative notion of "similarly situated" seems to override the methodological demands that racial contrasts vary in the full set of race perceptions. Researchers who claim to detect disparate treatment by isolating race perception from other so-called "non-race" perceptions must thus be explicit about the principle according to which they sort perceptions into the race versus the non-race perception bucket. If they are not so explicit, then this sorting appears post hoc. Effects are labeled "effects of race" only when inferencers take them to be wrongful, though they do not say on what normative theory they are wrongful.

We will argue in "The interpretive stakes of a constructivist view of race" section that there is nothing necessarily wrong with drawing on normative considerations in causal inference – indeed, researchers *must* draw on normative considerations to study discrimination. But researchers should be explicit about the underlying social and normative theory that guides their methodological choices. The problem is when researchers claim that what drives their sorting of effects into the race vs. confounder buckets is an empirical account of what is entailed in the treatment "perception of race," when what is really driving their methodology is an unstated normative theory about what kinds of race perceptions are discriminatory. All causal inferencers who claim to study the causal effects of race perception must put forward and defend a substantive assumption about what is entailed in the perception of race. Furthermore, researchers who do not define "disparate treatment" discrimination as coextensive with the causal effect of perception of race must explicitly express their normative theory of what constitutes discrimination. If this is not clear, we cannot understand why given empirical results do or do not count as evidence for the existence of that legal concept? The following section outlines what broad assumptions about race and racial perception are necessarily presupposed in such causal studies. Then we move to draw on CRT to fill in a substantive account of race and racial perception that we take to have significant theoretical and normative appeal.

Assumptions about race perception and their (lack of) justification

This section expands on two arguments introduced in the previous section. First, *any* study of a social category perception draws on – whether the researcher realizes it or not – a substantive account of what is entailed in a perception of that category. Second, such assumptions must be grounded in a theory of what kind of category it is in the society in question. To advance these points, it is essential to distinguish between four questions about race that are frequently conflated in discussions about its causal effects:

(1) What perceptions are entailed in the treatment *perception of race*?
(2) What *cues trigger* perception of race?
(3) What is necessary and/or sufficient to make someone a *member* of a particular "racial" category?
(4) What *is* race?

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

The first question asks about what mental states a researcher intends to bring about in treating a subject with "perception of race." The second asks how to bring about or trigger that perception. The third concerns the membership conditions (if any) of racial groups – i.e., what features of an individual make it apt, pursuant to a particular theory of "race," to classify them under a certain racial categorization? The fourth is a metaphysical question about the nature of the category of race – i.e., what kind of a category is "race"?

In our view, question (3) need not be a part of an investigation into causal inference about race because, as we explain below, a researcher can empirically study the cultural cognition of race and its effects while rejecting the culturally dominant conception of race membership or even while thinking there are no "races" on that definition (Wodak 2022). However, we argue in this section that answers to questions (1) and (2) are prerequisite to doing causal inference about race perception, and moreover, those assumptions must be grounded in a theory of (4): what race is.

Witches

Because race is a category about which many of us have extensive "prenotions" from living in a deeply racialized and racist society, it helps to illustrate our points with a fantastical example (Durkheim 2014: 39–46; Emirbayer and Desmond 2015:31–33). Consider a social category discussed at length in Karen and Barbara Fields' brilliant book *Racecraft: The Soul of Inequality in American Life*: witches (Fields and Fields 2012). Imagine a society where "witches" and "muggles" mark a salient and stratifying social category. Suppose that it is widely believed within this society that witches have occult supernatural powers and fly on brooms. Further, suppose that signifiers of witchhood include having warts on one's face and wearing a tall pointy hat. Nevertheless, in this culture (as depicted brilliantly in a *Monty Python* skit), pointy hats and warted noses are neither necessary nor sufficient for being a witch. Instead, the prevailing consensus is that what it takes for someone to be officially classified as a witch is that the person weighs less than a duck.[12] Many, but not all, members of this society believe that witches are born witches, some evil spirit enters the fetus causing mutation to witchhood and thus witchhood is seen as a natural (i.e., biological) category.

Ursula is a visiting anthropologist in this society interested in studying the causal effects of perceiving candidates with varying witchhood status ($D \in \{0,1\}$) who are doing "the same thing" under some – yet to be specified – description ($X = x$) on some outcome Y. She sets up an experiment in which she shows witchcraft cultural insiders a scene of a woman, who either has a pointy hat and warts on her nose or has a flat hat and no warts. In both scenes, the woman executes certain bodily movements: she moves her mouth and waves a small stick. Ursula then records some behavioral outcome of the subject who observes this scene: whether they do or do not scream ($Y \in \{0,1\}$).

Notice first that Ursula cannot even decide what stimuli or cues she should use in her study, until she defines *which* (excuse the pun) perceptions she seeks to bring about in her subjects. A commitment to the content of witchhood perception must precede the choice of instruments to stimulate that perception. She could, of course, choose to study something else, such as the causal effect of merely being exposed to assorted stimuli (like warts or pointy hats) but, to state the obvious, that is just a different

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

study. So long as Ursula aims to study what people in this society do when they *perceive witch/muggle status*, she must first identify the set of meanings, associations, stereotypes, etc. that constitute or make up the target perceptions.[13]

Clarifying the content of race perception is essential to specifying the counterfactual contrasts that comprise the causal effect of interest. For example, if Ursula assumes that the only perception she triggers in the minds of observers when she shows them the hat/nose cues are the beliefs that either "This person weighs less than a duck" or "This person weighs more than a duck," then she might think that the counterfactuals she is comparing are properly described in PO notation as: "Y_i (perception of stick movement and words, perception of witch) – Y_i (perception of stick movement and words, perception of muggle)." But if she assumes that those cues trigger a gestalt of meanings that operate as a schema through which the objective features of the action situation are given salience and perceptual content, then the counterfactuals she is comparing are better described as: "Y_i (perception of a witch casting a spell with a wand) – Y_i (perception of a muggle lady moving a branch and talking to herself)."

Furthermore, what Ursula posits as entailed in witch/muggle perception must be grounded in a theory of the witchhood category in this culture. Whether or not Ursula and her fellow researchers recognize this, causal analyses of witchhood perception invariably draw on an account of what kind of thing witchhood *is* in the society to form assumptions about what is entailed in the perception, and empirical evidence can be more or less consistent with that underlying account.[14] Knowledge about witchwood/muggle status in the society – both knowledge that it is a stratifying line and knowledge about how that stratification obtains – does more than motivate Ursula's research question. It also sets bounds on which assumptions about what cultural insiders perceive when they perceive witchwood/muggle status are plausible. Ursula's assumptions about the content of witchwood/muggle perception are subject to searching verification because the stereotypes, meanings, beliefs or associations activated by perception of witch cues and the conditions under which they are activated can be empirically studied. That is, she can test whether the assumption that the only thing triggered by apprehending a pointy hat and warty nose is the belief that "This person weighs less than a duck," as opposed to a perceptual schema.

Ursula can develop a theory of what perceptions are entailed in treating individuals with a perception of witchhood and what cues to use to do so (questions 1 and 2 above) by analyzing culturally dominant beliefs, associations and action. Her articulation of the culturally dominant content of witch/muggle perception does not mean that she herself endorses them as what is necessary and/or sufficient to make someone a witch or a muggle; nor is she thereby committed to a theory of what witches are (questions 3 and 4). Her ability to identify a consistent set of associations or meanings that are triggered when cultural insiders perceive witch cues – such as associations with occult powers and broom flying – is entirely compatible with her nonetheless believing that nobody in fact has those powers by virtue of weighing less than a duck (or nobody has those powers, period). She might personally reject the culturally dominant view of witches – the view that witchhood is a natural category existing by virtue of evil spirits – and, instead, take witchhood to be a socially constructed category that exists by virtue of collective meaning-making and material practices of controlling and denigrating non-conforming women.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

Race

Urusla's views on "witchhood" are akin to those that many scholars in the CRT tradition hold about race. Critical race theorists reject biological accounts of race in favor of a view that race is socially constructed (Delgado and Stefancic 2012: 21; Omi and Winant 1994: 64; Gómez 2010: 490; Haney-López 1994). Even while race is a biological fiction, certain "racial" traits that are related to biology – e.g., phenotypic and ancestry-based traits – may nonetheless trigger real causal effects because people in our society classify individuals into racial groups on the basis of those features (Gómez 2012: 231). Many scholars understand that these features define membership in racial categories in our society, even while they personally deny the socially predominant view that persons are members of racial groups by virtue of biological facts (Haney-López 1994: 7; Obasogie 2015: 3090).

In this section, we draw on CRT to argue for a particular theory of what race is. CRT teaches us that racial categories are categories defined by persons addressed by a set of meanings, material practices, social relations about shared genetic traits. Race is constructed out of these social facts. These facts constitute the grounds that endow so-called "racial" cues and traits (such as skin color, ancestry and certain names) with their significances.[15] Absent race-making social and historical processes, those cues and traits would not signify anything beyond themselves – that a person answers to a particular name or has a particular genetic trait. This, in turn, has implications for what kinds of assumptions about race perception are defensible and which are implausible. Having clarified the content of perceptions of race, we move to describe the race counterfactuals that compose common target causal estimands and accordingly, define the kinds of (un)equal treatment that, according to these studies, constitute discrimination.

What is race? Lessons from CRT

CRT sets forth a powerful theoretical framework that analyzes the relationship between race, racism and institutions of power, most notably the law. While CRT is a rich tradition that contains many different theoretical, methodological and normative commitments, scholars' analyses of race share some core tenets. In this section, we focus on two that bear on the central matter at hand in this paper.

One core tenet of CRT is that race is a social construction grounded in a set of social relations that constitute an unjust racial order (e.g., Delgado and Stefancic 2012: 8–9). A constructivist account of race posits that group distinctions exist only by virtue of an ongoing *process* of social construction (Omi and Winant 1994: 55–56). Critical race theorists furthermore remind us that race has an inherently *political* character. Racial categorizations emerge out of "power relations (subordination) and inequality (stratification)," which have their "historical roots in racial exclusion," and serve to "ideologically support[] a system of racial stratification" (Gómez 2010; 2012; Omi and Winant 1994: 55). Thus, many critical race theorists see the culturally dominant view of racial groups as "a natural division of human beings" (Obasogie et al. 2015: 3090) based on objective divisions in "morphology and/or ancestry" (Haney López 1994: 7) as false, nothing more than ideology. Nevertheless, race is "real" and causally efficacious because race is structurally embedded in virtually all major social institutions and racial meanings permeate social life as a system of thought and action.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

Second, the fact that race is defined by forms of social inequality explains a central tension in the concept of racial equality. Critical race theorists have elucidated how the law's official language of racial equality admits of many different conceptions of equality, some of which even work to entrench racial inequality (Bell 1992; Carbado 2022). The reason that there are multiple distinct conceptions of racial *equality*, with many of them being mutually exclusive, is precisely because race is a system of *inequality*: of subordination and domination, of the unequal distribution of social and material resources, and of differential evaluative meanings. Highly "formalistic," "restrictive" or "colorblind" notions of equality are compatible with the maintenance of a racially stratified order precisely because they turn a blind eye to the social facts that constitute race (Bonilla Silva 2009; Crenshaw 2019). By contrast, an "expansive" conception of equality aims at a racially equitable society, where this requires the "eradication of the substantive conditions of [racial] subordination" (Crenshaw 1988: 1341).

These two central insights of CRT – first, that race is grounded in a set of hierarchical social relations, and second, that there is an internal tension in the notion of equal treatment of groups defined by inequality – have important implications for empirical studies of the causal effect of race perception. First, race's social and political character means that racial meanings are neither natural or essential, nor are they entirely random. Rather, the content of racial perception derives from the racially stratified social structure. As Laura Gómez (2012: 231) writes, "To say that race is socially constructed is to acknowledge that we use phenotype or other visible characteristics to sort people into social groups [and] that we input qualities of good and bad to these groups." Gómez is here articulating racial classification as a schema that at once groups people in terms of their descriptive features (e.g., certain phenotypic features) and normative ones (e.g., certain evaluative notions of good and bad). These features are thereby entangled and co-constitute the category of race.

CRT leads us toward a *thick* view of what is entailed in perceptions of race. Even while racial classification might be based on physical features or ascriptions of ancestry, racial meanings consist in much more beyond these traits alone. As the above quote by Gómez suggests, someone who is treated with a perception that an individual is racialized a certain way is treated with a collection of associations, meanings, stereotypes and beliefs through which they apprehend, understand and evaluate that individual.

These theoretical premises about race and racial perception are also in line with "contemporary social psychological research [which] has exhaustively documented the fact that social groups can activate concepts," and more recently how "concepts (by themselves) can activate social groups" (Eberhardt et al. 2004: 876). Social psychological studies have shown that the meanings assigned to an individual's action alter as a function of the racial status that is ascribed to the individual. For example, Kunda and Thagard (1996: 286) found that the behavior of "pushing someone" is cast as "violent" when the individual involved is taken to be Black, whereas when the individual is white, he is less likely to be considered "aggressive" and as a result, the action is less likely to be construed as "violent." Moreover, ascriptions of racial status are inflected by perceptions of other contextual factors. Freeman et al. (2011: 7), for instance, found that individuals wearing high-status clothing were more likely to be categorized as white, whereas those wearing low-status clothing were more likely to be categorized as Black. Eberhardt et al. (2004: 877) hypothesize that the "bidirectionality" of influence

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

between social categories and concepts in perception function to tune attention to "relevant" aspects of some situation and thus are especially important to keep in mind when probing how agents' decision-making are guided by their perceptions.

CRT's account of how race is constituted in our society, taken together with empirical social psychological research on what cultural insiders in racially stratified societies perceive when they perceive race, shows that perceiving race is not a distinct, temporally removed event from perceiving the broader action situation. Rather, activating the category of "race" in an agent's mind entails activating a schema or lens through which they apprehend and give meaning to the entire action situation (Freeman et al. 2011).

The causal stakes of a constructivist view of race

This section shows the implications of embracing the constructivist view of race and the thick, cultural schema view of race perception. Our aim is to show that if one embraces these accounts of race and race perception, both the formal notation of causal inference and the dominant way that causal inferencers talk about the meaning of their target race-causal estimand is incomplete and, at times, even misleading. Contrary to what the formal notation and methodology of causal inference suggests, race perception may not be posited as fully distinct from the apprehension of other features of the action situation.

Let's return to the study discussed in "The "holy grail": isolating race perception" section, which seeks to isolate the causal effect of race on prosecutorial decisions. Equation (3) defined a standard race causal estimand that inferencers commonly target in such a study, the so-called "average causal effect" of race on prosecutorial decisions. We rewrite it here:

$$E\left[Y_i\left(D=b,\ X=x\right)-Y_i\left(D=w,\ X=x\right)\right] \tag{3}$$

Y indicates the prosecutorial outcomes; D indicates perception of the arrestee's race and X indicates perception of so-called "non-race" features of the arrestee and case. For example, some studies have used X to designate an arrestee's "behavior during a police encounter, their recorded criminal history, or both" (Gaebler et al. 2022: 28).

What are the counterfactual contrasts that define this causal effect? Consider, first, what it means on the assumption that perception of race entails only perception that the arrestee has a particular intrinsic trait (skin color, phenotype, genetic profile) or some ancestral fact. On this reading, Eq. (3) compares the charging decisions that a prosecutor makes in the following counterfactual contrasts: in one case, they perceive the arrestee to have b level of skin melanin (or b phenotype, or recent ancestors from b continent, etc.) and to have had zero prior arrests; in another case, they perceive the arrestee to have w level of skin melanin (or w phenotype, or ancestors from w continent, etc.) and to have had zero prior arrests.

By contrast, a constructivist view of race perception informed by core tenets of CRT theory takes it that treating decision-makers with perception of race entails treating them with a set of associations, beliefs, emotional or affective dispositions, and a schema through which various stimuli and information is given meaning. For example, suppose treating a prosecutor with a "perception of race" entails treating them

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

with a set of beliefs about social facts constituting the group to which the defendant is ascribed membership. Some of these social facts might give meaning to the information expressed in the *X* variable. Suppose the researcher assumes that perception of race entails, among other things, beliefs about an individual's relative risk of arrest. This assumption may be brought to the fore by explicitly denoting race perceptions "*w*" and "*b*" as vectors whose elements represent the various mental contents entailed in the perception:

$$w = \{w1, w2, \ldots \text{ lifetime risk of arrest lower than persons racialized Black}\}$$
$$b = \{b1, b2, \ldots \text{ lifetime risk of arrest higher than persons racialized white}\}.$$

If one adopts the constructivist assumption, the notation in Eq. (3) gives a misleading description of two counterfactual contrasts. For starters, it suggests that *D* and *X* are distinct causal factors. But the constructivist who sees racial perception as encompassing perception of many so-called "other" features of a situation will want to underscore that the prosecutor in this situation does not have two totally distinct perceptions: first, an arrestee who is racialized white, and second, an arrestee who has zero prior arrests. This way of putting it makes it seem like the prosecutor could be responding to the "zero prior arrests" aspect of an unracialized person in a way that is completely independent of the "racialized white" aspect. Rather, on a constructivist picture of race, the prosecutor faces up with the situation as a whole: a defendant, who is racialized white, has zero prior arrests.

The term "interaction" is sometimes used to describe an explanatory entanglement between *D* and *X*. But, as other scholars have also pointed out, "interaction" can be a misleading metaphor in this explanatory context because it implies that there are two distinct meanings that may interact to have some effect. For example, Taeku Lee (2008) urges us to abandon the dominant theoretical position within quantitative treatments of race which models "interactions" as two discrete, pre-existing explanatory entities coming together or mixing. He urges us to embrace a relational understanding of "interactions" which views such variables as picking out the very processes by which (e.g.) race is politicized and politics is racialized. Lee quotes Mustafa Emirbayer (1997), who writes that "attempts by statistical researchers to 'control for third variables …' ignore the ontological embeddedness or locatedness of entities within actual situational contexts" (289; see also Dembroff 2023).

The constructivist view on offer here is in line with these authors' contention: it is not that *X* and *D* "interact," but that there is no standalone interpretation of the *X* features. Just as it is inapt to say that the effect of reading the letters "effect" is a matter of the effect of reading the letter "e" *interacting* with the effect of reading the letters "ffect," it is here inapt to say that the effect of perceiving a person to be racialized "interacts with" the effect of perceiving a person who has zero priors. The *X* features pertain *to the racialized person* – e.g., a particular person was perceived to have behaved in a specific way with the police (or was recorded to have done so in the dataset), a particular person had a specific history of arrests and conviction – and that person was racialized in a specific way. The *X* features are not free-floating descriptors that describe just anyone or no one at all. The *purpose* of these studies is to ascertain whether decision-makers respond

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

differently to credentials when those credentials pertain to differently racialized candidates. For these reasons, the better way of expressing the target estimand in these studies is:

$$E\left[Y_i\left(D = b\,DX = bx\right) - Y_i\left(D = w\,DX = wx\right)\right] \tag{4}$$

We understand that this notation is nonstandard, but the constructivist account of race calls for it for two reasons. First, the new variable DX gives expression to the whole racialized criminalized person; it is *not* a compound term in the sense of being the product, in a mathematical sense, of two separable features of a person, D and X. (Recall the analogy to the word "effect": the effect of "effect" is of the *whole word*, not of a compound treatment of individual letters.) Denoting it in this way encodes the constructivist view that two persons who have the same number of prior arrests and so share X values are not necessarily thereby the "same" in all respects but for racialized status. This is because the beliefs, associations and meanings entailed in D are brought to bear on apprehending, interpreting and acting on X. If X has its causal powers only by virtue of being interpreted in light of D, then the two differently racialized arrestees could also have differently causally efficacious arrest records. The variable DX, with wx and bx as values of DX, reflects this theoretical stance. Second, there is no comma between D and DX in Eq. (4) indicating a bundled treatment. This notation expresses that the *point* of these studies is to make the decision-maker believe both that the candidate has a particular racialized status *and* that the criminal history or credentials listed in the file pertains to someone who has that racialized status (b not w). Said another way, subjects would be non-compliant with treatment if, for example, they thought that the information in the file (the X features) pertained to someone other than the person that was racialized (e.g., $D = b$).

Equation (4) compares the charging decisions that a prosecutor makes in the following counterfactual contrasts: in one case, they perceive racialized black social meanings {$b1$, $b2$, … lifetime risk of arrest higher than for persons racialized white} and that this defendant who (inter alia) has a higher lifetime risk of arrest than persons racialized white has had zero prior arrests. In the other case, they perceive racialized white social meanings {$w1$, $w2$, … lifetime risk of arrest lower than for persons racialized Black} and that this candidate who (inter alia) has a lower lifetime risk of arrest than persons racialized Black has had zero prior arrests. This difference matters immensely for how causal inference studies about race perception are interpreted. We will draw on a recent study to show precisely how.

The interpretive stakes of a constructivist view of race

This section picks up our running example to illustrate the stakes of adopting the constructivist view. In "A Causal Framework for Observational Studies of Discrimination," Gaebler et al. (2022: 39) find that Black and white arrest files, conditional on sharing the same arrest charges, number of prior arrests recorded on the rap sheet and so on, are charged at similar rates.[16] The authors interpret the results of their causal inference exercise to be evidence of a legal-normative state of affairs. Specifically, they see their "empirical findings as providing evidence that perceived gender and race have limited effects on prosecutorial charging decisions in the jurisdiction we considered." However, they note two caveats to this conclusion. First, they write that the estimand they measure captures "discrimination in the charging decision, and, *in particular*, is

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

not designed to capture the *cumulative* effects of discrimination stemming from arrests and other earlier decision points" (39; emphasis added). Second, they note that their estimate might be subject to unmeasured confounding (though they say that their sensitivity analysis shows risk of such confounding to be slight) (39).

Do Gaebler *et al.*'s causal analysis indeed licenses the determination that there is little discrimination in the prosecutorial charging decisions in their studied jurisdiction? We want to draw out two places in their analysis where substantive assumptions, unacknowledged in the text, powerfully drive their conclusions. Specifically, Gaebler *et al.*'s caveats rely on two sets of distinctions: first, between cumulative and prosecutorial-specific racial discrimination, and second, between the causal effect of race perception "itself' and confounders of it. As we will show, both distinctions are only tenable on a non-constructivist view of race and race perception. Before moving on to discuss Gaebler *et al.*'s work, we should emphasize that these assumptions are widespread in this body of work and, in our view, confuse questions of sociology and statistics, morals and methods. Our aim in dissecting their analysis is not to pick out a distinctively problematic case of these assumptions but rather to use this clearly written piece to dialogue with views that are standard in the literature.

The social meaning of race entails its cumulative effects

Gaebler *et al.*'s first caveat is that their results do not reflect "the cumulative effects of discrimination stemming from arrests and other earlier decision points"; instead, they clarify that in their view the target estimand only captures "discrimination in the second-stage charging decision" (28). They equate this quantity with disparate treatment *by prosecutors*, as opposed to the accumulated disadvantages that arrestees bring with them from prior stages of life or earlier in the criminal process. This distinction between a differential due to discrimination by some bounded decision-makers and a differential due to the "cumulative" process of disadvantage is ubiquitous in the law and in social scientific work on detecting discrimination (e.g., *Washington v. Davis at 242; Village of Arlington Heights v. Metropolitan Housing Dev. Corp.* at 265). Still, we want to challenge the logic of this dominant view.

The authors claim that "cumulative effects of discrimination stemming from arrests and other earlier decision points" (39) explain only "selection into the sample of interest" (26). In other words, prior discrimination bears only on the likelihood that a given racialized person will be arrested. The implicit picture here is that progression through the criminal legal process is the result of chance mechanisms such as weighted coin tosses, wherein being racialized amounts to nothing more than being burdened or blessed with a certain probability of progressing to subsequent stages (e.g., being stopped, arrested, charged, etc.). But these cumulative effects of discrimination have no implications for the content of race perception, beyond its effect on the racial composition of the population of people who are arrested.

In our view, this sharp distinction between the "cumulative effects of discrimination" and "prosecutorial-specific discrimination" is deeply misguided. A constructivist insists that the cumulative discrimination bears not just on the racial composition at various stages of the criminal legal system but also on the cognitive content of race perception. Race perception has the content it does precisely *because of* the "cumulative effects of discrimination stemming from arrests and other earlier decision points"

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

(Gaebler et al. 2022: 39). In our world, progression to subsequent stages in the criminal legal process is not determined by chance mechanisms but by human decision-makers who act on the basis of reasons – ostensibly, on the basis of legal criteria. These decision-makers may not make decisions well, carefully, or by faithfully adhering to the legal criteria specified for each decision point. Indeed, the very concern with discrimination in this domain is that such actors are making decisions in a racially unfair and unjust way under cover of an ideology of impartial legality. So, the social meanings that accrete to the stratifying traits in our society – skin color, phenotype, real or perceived ancestry – are generated by the beliefs and actions of other actors. As W. E. B. Du Bois (1903: 14) and Khalil Gibran Muhammad (2011) have forcefully argued, the social meaning of Blackness comes from cultural beliefs that persons racialized Black are overrepresented in the criminal legal system not because of bad luck but because of bad behavior.

Our point is not that these authors *should* have measured the "cumulative discrimination stemming from both the arrest and charging decisions" (32). Rather, it is that the distinction between cumulative discrimination from prior stages and discrimination at a particular stage assumes a thin conception of race perception. For a constructivist, there is no way to eliminate the effect of "cumulative discrimination" with statistical or experimental procedures, because the social meaning of race that is perceived by prosecutors entails the meanings that have accreted from these iterative interactions.[17] As Angela Onwuachi-Willig and Mario Barnes put it, "it is not physical race but … rather the constructed social meanings of race [] that trigger both conscious and unconscious forms of discrimination" (2005: 6). These scholars argue that persons can be subject to racial discrimination irrespective of their "true" (according to socially dominant definitions) category membership because the wrongfulness of the act lies in responding to these nefarious social meanings and thereby remaking them (2005: 20).

If perceiving race entails picking up meanings accumulated from processes of disadvantage, then "cumulative discrimination" is not a confounder of the causal effect of perceiving race, rather it is partly constitutive of it. Accordingly, if the stated aim is to make use of such causal studies to detect racial discrimination, inferencers must take a stand on what treatment is fair or just *in light of* those meanings. This brings us to Gaebler *et al.*'s second caveat.

Deconfounding as making normative, not statistical, assumptions

Recall that Gaebler *et al.* – like many causal inferencers working in this vein – simultaneously make two commitments: a commitment to identifying the causal effect of prosecutors' perception of defendants' race on charging decisions[18] and a commitment to measuring "disparate treatment" racial discrimination.[19] These inferencers sometimes suggest that these two commitments are aligned: stating that empirical "evidence that perceived gender and race have limited effects on prosecutorial charging decisions," subject to unmeasured confounding, is simultaneously dispositive of whether there is "disparate treatment" discrimination.[20] On this view, the legal-normative designation cannot follow without prior sociological assumptions about what is entailed in race perception. This theory is what adjudicates between something being part of race perception versus a confounder of it.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

On a thin view of race perception, the full extent of what prosecutors perceive is something like (e.g.) skin or ancestry. So, everything that is not (e.g.) skin or ancestry perception, counts as a potential confounder. The claim that they successfully "deconfound" the effect of race perception only when they present the decision-makers with the identical formal stimuli – e.g., same number of prior arrest, same arrest location – rests on the assumption that providing this information makes the candidates *more* similar, not less, in the eyes of the decision-makers, lest it be a "bad control" (Angrist and Pischke 2009).

The constructivist view of race perception suggests a different approach to what it means to "deconfound" in this empirical exercise. For the constructivist, different race perception consists in a set of differential expectations, opportunities and meaning frameworks. Accordingly, "deconfounding" must be driven by normative assumptions about which racial perceptions or beliefs are fair and just to act on. The constructivist takes it that there is no way to make the differently racialized arrestees identical in *all* respects, because different race perception consists in more than perception of thin traits such as skin or ancestry. On this view, attempts to make arrestees more similar in one respect – e.g., by making them have the same number of priors or have the same arrest location – necessarily make them different in other respects. To see why, return to the estimand in Eq. (4). That notation takes care not to express the assumption that arrestees that have formally identical X features are "the same" in a substantive or causally relevant sense. In one sense, the two arrestees are perceived to be the same – e.g., the prosecutor reads that they share the same number of prior arrests. But in another sense, they are perceived to be different. For example, the arrestee who is racialized Black is perceived to have an "unexpectedly" low number of arrests relative to the risks they face by virtue of being racialized Black in our society, whereas the candidate who is racialized white has an "expected" number of prior arrests relative to the risks they face by virtue of being racialized white in our society. Because race is a marker of social inequality that differentially positions individuals along a number of dimensions, making some so-called "non-race" features the same will, necessarily, make others different (Kohler-Hausmann 2024; Hu, forthcoming). It bears recalling that central insight of critical race theorists about "colorblind" or "formal" equality: treating people "equally" who have been treated unequally is not a race-neutral theory of justice, it is race-conscious theory of justice that takes an affirmative stand on what allotting the "same" treatment demands in the face of these inequalities (Bell 1992; Greene 1990; Crenshaw).

So, an analyst must pick *which* features they think must be made the same between the differently radicalized candidates (i.e., define the relevant similarities that merit equal treatment). The constructivist recognizes that picking the relevant "similarity" is an irreducibly normative exercise. What counts as a "confounder" in discrimination detection requires sociological and normative assumptions, both of which must precede methodological determinations. Consider the case of arrest location. A constructivist might posit that expectations about what neighborhoods people frequent are part of racial perception. Imagine a study that tells prosecutors that two arrestees – one racialized Black and the other white – were both arrested at the same arrest location, say, Brownsville, Brooklyn (a low-income, largely Black neighborhood). Telling a prosecutor that both arrestees were arrested in Brownsville, Brooklyn does not void or strike out the aspect of racial perception about neighborhood residence.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

Rather, it makes the white arrestee "unexpected" at the arrest location and the Black arrestee "expected" at the arrest location. Giving the differently racialized arrestees the same arrest location makes them the same in one respect but different in another.

Once we see that presenting differently racialized candidates as having identical formal features does not in fact substantively equalize them in all causally relevant ways, we can open up causal studies for re-interpretation. Our suggestion is that such studies in fact express the conditions under which equal treatment is expected *according to a particular normative theory*. Thus, an observed difference in prosecutorial decisions made between differently racialized persons with the same number of arrests or different arrest locations counts as discriminatory only if prosecutors *ought* to treat these individuals similarly when (e.g.) they have the same number of prior arrests or the same arrest locations, notwithstanding the fact that prosecutors might assign different substantive meanings to those facts in light of race. Putting it this way brings out a normative assumption without which no conclusion regarding racial discrimination can follow.

Making these assumptions explicit is critical for interpreting the results of these studies. For example, suppose a study shows a *higher* charging rate for white arrestees compared to Black arrestees even after prosecutors have been given formally identical information about the arrestees. Does this demonstrate discrimination against white defendants? In our view, this question cannot be answered unless we commit to a view regarding how prosecutors *ought* to treat differently racialized arrestees *in light of* the social meaning of race. If, for example, one thinks that it is not discriminatory for prosecutors to draw on their background knowledge about the lifetime risks of arrest faced by persons racialized Black to give substantive meaning to the number of prior arrests because of discriminatory policing, living in neighborhoods of concentrated poverty and countless other forms of accumulated disadvantage, then evidence of a lower charge rate for arrestees racialized white compared to arrestees racialized Black with the same number of prior arrests is not evidence of racial discrimination.

Conclusion

The law and society movement has long thought about how scholarship engages social and legal change. This article urges a dramatic rethinking of causal inference about race perception. Some readers might be concerned that our argument threatens to undercut a body of work that powerfully presents evidence *for* the pervasiveness of racial discrimination. After all, research documenting the causal effects of race has been not only scientifically valuable but politically and legally useful in the fight for racial justice. Work by scholars such as Pager and Quillian (2005); Bertrand and Mullainathan (2004) and Kline et al. (2022) reveal the prevenance of discrimination. Such scholarship can play a crucial role in agitating for a more just future by exposing the great moral and political shortcomings of our collective life.

We want to be clear that we do not impugn the normative conclusions drawn from much of this work. But, as we have argued, the grounds for those conclusions are not based solely in sound methodological practice. Rather, they issue from a combination of theoretical, empirical and, importantly, *normative* propositions that set forth a thoroughly value-laden conception of racial discrimination. In our view, it is shortsighted

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

to avoid interrogating the assumptions upon which these studies rest on account of a set of favorable findings. To say that these studies prove race discrimination because of their methodological rigor is to agree to much more than a set of outcomes. It is to cede to an entire form of reasoning: that racial discrimination is to be defined as encompassing *all* and *only those* deviations from some standard of equality that is based on an unreflectively drawn distinction between "race" and "non-race."

Such a concession is a dangerous gambit. Not all empirical studies of discrimination will bolster progressive positions on social justice. The sparring expert reports in *SFFA* are an object lesson in how fickle arguing on these terms can be when the key assumptions are hidden from view. Peter Arcidiacono, the expert for SFFA, and David Card, the expert for Harvard spoke freely of racial "tips" or "penalties" with no mention of the normatively laden baseline against which such deviations must be measured. Setting aside the specifics of those analyses – about which we have much to say but must leave for another day – this framing hides the fact that all inferences from statistical claims to legal-normative ones require assumptions. In this case, both analyses operated from an impoverished theory of race and unstated views of *what* equality is owed in light of race. Accepting these lines of debate as they are drawn means foregoing broader arguments about when non-discrimination not only *allows* causal effects of race but in fact *demands* them.

Historians and sociologists of science have used the term "mechanical objectivity" to describe an ideal of inquiry that seeks to arrive at conclusions solely through applying strict and explicit rules and standards, thereby minimizing the need for exercises of interpretation, judgment or discretion (Daston and Galison 1992; Espeland 1997; Porter 1996). The drive to reduce analyses of racial discrimination to exercises of causal inference reflects this ideal. And importantly, the rise of statistical and causal inference-based evidence in discrimination cases indicates that the law, too, sometimes claims to draw verdicts about discrimination from mechanical rule-based exercises, ostensibly free of the taint of human subjectivity.

Law and society scholars have studied the processes by which the law variously bolsters, contests and itself traffics in the ideal of mechanical objectivity as a part of this broader strategy of legitimation. As these scholars continually remind us, "mechanical objectivity can never be purely mechanical" (Porter 1996: 5). This is clear as day in the case of causal inference analyses about discrimination, which encompass statistical, mathematical and logical reasoning *as well as* sociological and normative reasoning. The latter forms of reasoning are what generate the model of the world to which formal methods are then applied. Nothing internal to statistics or causal inference determines whether a given mathematical model is an adequate representation of how the world works, or whether the variables defined within the model tracks an explanatorily fruitful social ontology. These sociological and normative premises are irreducibly evaluative. They necessarily contain human interpretation and judgment, which must themselves be defended. Yet, such premises are often treated as though they have been simply been given to us ready-made, and conclusions based on these premises are often treated as they are deductively true as a matter of logic internal to the technical fields.

The discipline of law and society is therefore well placed to interrogate the ramifications of the gradual redefinition of the terrain upon which battles over discrimination are won or lost. When causal inference takes center stage as the arbiter of such cases,

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

discrimination as a legal and moral concept is held hostage to technical matters of methodology. Debate about discrimination must now always route through a specialized form of causal and statistical reasoning (e.g., *United States v. Johnson* 2015; *Floyd v. City of New York* 2013; *United States v. Duque-Nava* 2004; *United States v. Jones* 1999; *United States v. Payne* 2015). Worse still, inferencers are often rewarded for hiding or obfuscating which theory of race and discrimination they are operationalizing. This is what we see in many of the current debates among experts, which are framed as disagreements about statistics not sociology, methods not morals (see, e.g., Brief for Economists as Amicus Curiae in Support of Petitioners 2021: 4–5; Expert Report of Peter Arcidiacono 2016: 7–21; Durlauf and Heckman 2020; Knox et al. 2020; Gaebler et al. 2022). Meanwhile, arguments that cannot be recast in these terms are ruled as out of bounds, inadmissible or, even worse, irrelevant. In a regime in which disputes over discrimination reduce to exercises in causal detection, expertise in statistics converts into expertise on normative matters, or more perniciously, threatens to eclipse entirely discrimination's fundamentally moral and political character.

Acknowledgements. We thank Robin Dembroff, Daniel Malinsky, Tracey Meares, Frederick Eberhardt and Gideon Yaffe for their generous feedback. Lily Hu is grateful for support from the Jain Family Institute. Issa Kohler-Hausmann is grateful for support from the Simons Center for Theoretical Computing.

We thank P. Aronow, Mario Barnes, Robin Dembroff, Frederick Eberhardt, Daniel Malinsky, Tracey Meares, Osagie Obasogie, Fredrik Sävje, Gideon Yaffe, and the editors and three anonymous reviewers at *Law & Society Review* for their helpful comments.

Conflict of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Notes

1. E.g., *Comcast Corp. v. Nat'l Ass'n Afr. Am.-Owned Media,* 140 S. Ct. 1009, 1019 (2020) ("To prevail, a plaintiff must initially plead and ultimately prove that, but-for race, it would not have suffered the loss of a legally protected right."). National Research Council's report "Measuring Racial Discrimination," in a chapter entitled "Causal Inference and the Assessment of Racial Discrimination," states 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?" (2004: 77). In other work, we have argued that this causal definition of discrimination is wrong as a theoretical and legal-interpretative matter (Kohler-Hausmann 2018; Hu & Kohler-Hausmann 2020; Dembroff & Kohler Hausmann 2022), but this article accepts that definition in order to interrogate what precisely it means.

2. According to Nobel Prize-winning economist James Heckman, "Discrimination is a causal effect defined by a hypothetical *ceteris paribus* conceptual experiment-varying race but keeping all else constant" (1998: 102). See also Kline et al. (2022: 7–8) and Starr (2016: 485–88).

3. "Estimates of disparate treatment discrimination are estimates of causal effects, not mere correlations – specifically, the causal effect of citizens' race (or of the racial compositions of communities) on police decision-making" (Starr 2016: 501).

4. SFFA, 600 U.S. 181 at 298 (Gorsuch J., concurring). *See also United States v. Johnson* 2015; *Floyd v. City of New York* 2013.

5. Our claims apply in equal measure to other schools of causal inference such as the Structural Causal Modeling approach, which uses directed acyclic graphs. Despite many disagreements across these frameworks, they share commonalities sufficient to ground the discussion here.

6. The treatments of "perception" vs. "exposure" are different treatments; they correspond to different study designs and lead to different inferences and study conclusions. The researchers that we engage with in this article characterize their studies, methods and discuss their findings in ways that make clear they use the term "race perception" to mean that a decision-maker is treated with cultural category

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

cognition or formation of racial beliefs, not is not merely exposed to some stimuli. E.g., Crabtree et al. 2023: 2; Bertrand and Mullainathan 2004: 991.

[7] We use "racialized" throughout the paper for the reasons explained by many critical race theorists, that race is not an intrinsic trait people possess but a relational property one has by "living as a 'raced' person" (Onwuachi-Willig and Barnes 2005: 19; see also Gotanda (2000: 1694)).

[8] Typically, notation inside the parentheses refers to the treatment, and the "all else equal" idea is expressed by noting that the same unit i receives both treatments. We include "$X = x$" inside the parentheses to make explicit that the point of the study design is to create the perception that each candidate has the racial status listed in the file *and* the credentials listed in the file.

[9] "[T]he 'race effect' for individual i is $\tau = Y_1 - Y_0$ – that is, the difference in Y that can be attributed to an individual's race. This quantity is the proverbial 'holy grail' – the parameter that we are all attempting to estimate but never quite do" (Fryer 2018: 2).

[10] Other scholars have noted that observational studies documenting racial disparities that "control for" many factors seem to also have causal aims, though they are often not forthcoming about those aims (e.g., Lundberg et al. 2021; Knox et al. 2020a; Grossman et al. 2023: 94).

[11] These inferences seem to embrace a mental state view of "disparate treatment" – that what makes an act discriminatory is that the decision-maker was guided by some prohibited mental state of acting "on the basis of race." But, like the law, they are not clear if they mean that all racial mental states are discriminatory, or only some normatively defined subset are (Kohler-Hausmann 2024).

[12] In *Monty Python and the Holy Grail*, a mob of villagers hauls a woman dressed up with a carrot on her nose and a funnel on her head to the authorities demanding to "burn her!" as a witch. The priest (or knight?) explains that, despite these cues that indicate witchhood, the real test of whether someone is a witch is if she weighs more than a duck. The scene shows that the necessary and sufficient conditions for membership in a social group may not overlap with the cues that trigger the meanings and associations of the social group.

[13] Many race-causal studies proceed to cue selection without making clear their assumptions with respect to what those cues are supposed to trigger. If their aim is to simply study the effects of stimuli presentation, then calling the study one of "race perception" is unwarranted.

[14] Charles Mills illustrates the difference between a thick marker of difference and a thin distinction in his discussion of "race" vs. "quace" (Mills 1998: 42).

[15] Different cues may trigger different cognitive content about race. A decision-maker could be treated with a file containing a race checkbox, a description of a person, or might visually apprehend a racialized person. When a causal inferencer groups all these treatments together as treatments of "perception of race," they assume that these different cues all trigger the intended perceptual content, despite the differences in the precise cognitive content triggered by each.

[16] Another study shows that, conditional on sharing the same arrest charges, number of prior arrests recorded on the rap sheet and various other covariates, white misdemeanor arrestees are charged at a *higher* rate than black arrestees (Kohler-Hausmann 2022). The standard account would count this as evidence of discrimination against white misdemeanor arrestees.

[17] Said in technical terms, the statistical assumption of *overlap* or *positivity* can "solve" the selection part of the problem but cannot "solve" the fact that race perception entails beliefs and expectations about who is expected to be selected into certain stages.

[18] "The estimand in Equation (1) compares the potential second stage decisions under two race perception scenarios. For example, it compares the potential charging decisions when the prosecutor perceives the individual to be either Black or White" (29).

[19] A "central aim of this article is to formalize technical assumptions that allow one to statistically identify discrimination – more precisely, disparate treatment – in the second stage (e.g., in prosecutorial charging decisions)" (277). For other places where the authors state their causal quantity is identical to "disparate treatment," see also pp. 26–28, 31–33, 37–39.

[20] The authors equate the fact that "perceived [] race" shows limited effects subject to unmeasured confounding with the effect of "disparate treatment," writing: "The second-stage sample average treatment effect [] captures discrimination in the second-stage decision among those who made it past the first stage (e.g., discrimination in charging decisions among those who were arrested). This estimand maps onto a common understanding of second-stage decisions, including in our charging example" (28).

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

References

Agan, Amanda and Sonja Starr. 2018. "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment." *Quarterly J. of Economics* 133 (1): 191–235. doi:10.1093/qje/qjx028.

Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Bell, Derrick. 1992. "Racial Realism." *Connecticut Law Rev.* 24 (2): 363–79.

Bertrand, Marianne and Esther Duflo. 2017. "Field Experiments on Discrimination." *Handbook of Economic Field Experiments* 1: 309–93.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Rev.* 94 (4): 991–1013. doi:10.1257/0002828042002561.

Block, Ray, Charles Crabtree, John B. Holbein and J. Quinn Monson. 2021. "Are Americans Less Likely to Reply to Emails from Black People Relative to White People?" *Proceedings of the National Academy of Sciences.* 118: e2110347118.

Bonilla-Silva, Eduard. 2009. *Racism without Racists: Color-Blind Racism and Racial Inequality in Contemporary America*. New York: Rowman and Littlefield.

Carbado, Devon. 2022. "Strict Scrutiny and the Black Body." *UCLA Law Rev.* 69 (2): 2–77.

Crenshaw, Kimberlé. 1988. "Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law." *Harvard Law Rev.* 101 (7): 1331–87. doi:10.2307/1341398.

Crenshaw, Kimberlé.2019. In Unmasking Colorblindness in the Law: Lessons from the Formation of Critical Race Theory. *Seeing Race Again: Countering Colorblindness across the Disciplines*, edited by Kimberlé Crenshaw, Luke Harris, Daniel HoSang and George Lipsitz, 52–84. Oakland: The Regents of the University of California.

Daston, Lorraine and Peter Galison. 1992. "The Image of Objectivity." *Representations* 40: 81–128. doi:10.2307/2928741.

Delgado, Richard and Jean Stefancic. 2012. *Critical Race Theory: An Introduction*. New York: NYU Press.

Dembroff, Robin. 2023. "Intersection Is Not Identity, or How to Distinguish Overlapping Systems of Injustice." In *Conversations in Philosophy, Law, and Politics*, edited by Ruth Chang, and Amia Srinivasan. New York, USA: Oxford University Press 383-398.

Doleac, Jennifer L. and Benjamin Hansen. 2016. *Does 'Ban the Box' Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories are Hidden*. Cambridge, MA: National Bureau of Economic Research.

Du Bois, W.E.B. 1903. *The Souls of Black Folk*. Chicago: A. C. McClurg & Co.

Durkheim, Emile. 2014. *The Rules of Sociological Method: And Selected Texts on Sociology and Its Method*. New York: Simon and Schuster.

Durlauf, Steven N., and James J. Heckman. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment." *J. of Political Economy* 128 (10): 3998–4002. doi:10.1086/710976.

Eberhardt, Jennifer L., Phillip Atiba Goff, Valerie J. Purdie and Paul G. Davies. 2004. "Seeing Black: Race, Crime, and Visual Processing." *J. of Personality and Social Psychology* 87 (6): 876–93. doi:10.1037/0022-3514.87.6.876.

Emirbayer, Mustafa. 1997. "Manifesto for a Relational Sociology." *American J. of Sociology* 103 (2): 281–317. doi:10.1086/231209.

Emirbayer, Mustafa and Matthew Desmond. 2015. *The Racial Order*. Chicago: University of Chicago Press.

Espeland, Wendy Nelson. 1997. "Authority by the Numbers: Porter on Quantification, Discretion, and the Legitimation of Expertise." *Law & Social Inquiry* 22 (4): 1107–33. doi:10.1111/j.1747-4469.1997.tb01100.x.

Fields, Karen and Barbara J. Fields. 2012. *Racecraft: The Soul of Inequality in American Life*. New York: Verso Books.

Freeman, Jonathan B., Andrew M. Penner, Aliya Saperstein, Matthias Scheutz and Nalini Ambady. 2011. "Looking the Part: Social Status Cues Shape Race Perception." *PLoS One* 6 (9): e25107. doi:10.1371/journal.pone.0025107.

Fryer, Roland. 2018. "Reconciling Results on Racial Differences in Police Shootings." *AEA Papers and Proceedings.* 108: 228–33.

Fryer, Roland and Steven Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly J. of Economics* 119 (3): 767–805. doi:10.1162/0033553041502180.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

Gaddis, S. Michael. 2017. “How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies.” *Sociological Science* 4 (19): 469–89. doi:10.15195/v4.a19.

Gaebler, Johann, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel and Jennifer Hill. 2022. “A Causal Framework for Observational Studies of Discrimination.” *Statistics and Public Policy* 9 (1): 26–48. doi:10.1080/2330443X.2021.2024778.

Glymour, Clark. 1986. “Statistics and Causal Inference: Comment: Statistics and Metaphysics Journal of the American Statistical Association.” 81 (369): 964–66.

Glymour, Clark and Madelyn R. Glymour. 2014. “Race and Sex are Causes.” *Epidemiology* 25 (4): 488–90. doi:10.1097/EDE.0000000000000122.

Gómez, Laura E. 2010. “Understanding Law and Race as Mutually Constitutive: An Investigation to Explore an Emerging Field.” *Annual Rev. Law and Social Science* 6: 487–505. doi:10.1146/annurev.lawsocsci.093008.131508.

Gómez, Laura E. 2012. “Looking for Race in All the Wrong Places.” *Law & Society Rev.* 46 (2): 221–45. doi:10.1111/j.1540-5893.2012.00486.x.

Gotanda, Neil. 2000. “Comparative Racialization: Racial Profiling and the Case of Wen Ho Lee.” *UCLA Law Rev.* 47 (4): 1689–703.

Greene, Linda. 1990. “Race in the Twenty-First Century: Equality through the Law?” *Tulane Law Rev.* 64 (6): 1515–41.

Greiner, D. James and B. Rubin Donald. 2011. “Causal Effects of Perceived Immutable Characteristics.” *The Rev. Economics and Statistics* 93 (3): 775–85.

Greiner, James D. 2008. “Causal Inference in Civil Rights Litigation.” *Harvard Law Rev.* 122 (2): 533–98.

Grossman, Joshua, Julian Nyarko and Sharad Goel. 2023. “Racial Bias as a Multi-stage, Multi-actor Problem: An Analysis of Pretrial Detention.” *J. of Empirical Legal Studies* 20 (1): 1–48. doi:10.1111/jels.12343.

Guryan, Jonathan and Kerwin Kofi Charles. 2013. “Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots.” *The Economic J.* 123 (572): 417–32. doi:10.1111/ecoj.12080.

Haney-López, Ian F. 1994. “The Social Construction of Race: Some Observations on Illusion, Fabrication, and Choice.” *Harvard Civil Rights-Civil Liberties Law Rev.* 29 (Winter): 1–62.

Heckman, James J. 1998. “Detecting Discrimination.” *J. of Economic Perspectives* 12 (2): 101–16. doi:10.1257/jep.12.2.101.

Heckman, James J. 2005. “The Scientific Model of Causality.” *Sociological Methodology* 35: 1–97. doi:10.1111/j.0081-1750.2006.00164.x.

Ho, Daniel E. and Donald Rubin. 2011. “Credible Causal Inference for Empirical Legal Studies.” *Annual Rev. Law and Social Science* 7: 7–40. doi:10.1146/annurev-lawsocsci-102510-105423.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *J. of American Statistical Association* 81 (396): 945–60. doi:10.1080/01621459.1986.10478354.

Holzer, Harry J., Steven Raphael and Michael A. Stoll. 2006. “Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers.” *J. of Law and Economics* 49 (2): 451–80.

Hu, Lily. forthcoming. “Normative Facts and Causal Structure.” *J. of Philosophy*.

Imbens, Guido, and Donald Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Kline, Patrick, Evan K. Rose and Christopher R. Walters. 2022. “Systemic Discrimination Among Large U.S. Employers.” *The Quarterly J. of Economics* 137 (4): 1963–2036. doi:10.1093/qje/qjac024.

Knox, Dean, Will Lowe and Johnathan Mummolo. 2020a. “Administrative Records Mask Racially Biased Policing.” *American Political Science Rev.* 114 (3): 619–37. doi:10.1017/S0003055420000039.

Knox, Dean, Will Lowe, and Johnathan Mummolo. 2020. “Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection?”.

Kohler-Hausmann, Issa. 2022. “Don’t Call It a Comeback: The Criminological and Sociological Study of Subfelonies.” *Annual Rev. Criminology* 5: 229–53. doi:10.1146/annurev-criminol-070221-024802.

Kunda, Ziva and Paul Thagard. 1996. “Forming Impressions from Stereotypes, Traits, and Behaviors: A Parallel-constraint-satisfaction Theory.” *Psychological Rev.* 103 (2): 284–308. doi:10.1037/0033-295X.103.2.284.

Lee, Taeku. 2008. “Race, Immigration, and the Identity-to-Politics Link.” *Annual Rev. Political Science* 11: 457–78. doi:10.1146/annurev.polisci.11.051707.122615.

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

Lundberg, Ian, Rebecca Johnson and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Rev.* 86 (3): 532–65. doi:10.1177/00031224211004187.

Mills, Charles. 1998. *Blackness Visible*. Ithaca: Cornell University Press.

Muhammad, Khalil G. 2011. *The Condemnation of Blackness*. Cambridge: Harvard University Press.

National Research Council. 2004. *Measuring Racial Discrimination*. Washington, DC: The National Academies Press.

Obasogie, Osagie K., Julie N. Harris-Wai, Katherine Darling and Carolyn Keagy. 2015. "Race in the Life Sciences: An Empirical Assessment, 1960-2000." *Fordham Law Rev.* 83 (6): 3089–114.

Omi, Michael and Howard Winant. 1994. *Racial Formation in the United States: From the 1960s to the 1990s*. New York: Routledge.

Onwuachi-Willig, Angela and Mario L. Barnes. 2005. "By Any Other Name: On Being Regarded as Black, and Why Title VII Should Apply Even if Lakisha and Jamal are White." *Wisconsin Law Rev.* 2005 (5): 1283–343.

Pager, Devah and Lincoln Quillian. 2005. "Walking the Talk? What Employers Say versus What They Do." *American Sociological Rev.* 70 (3): 355–80. doi:10.1177/000312240507000301.

Pager, Devah and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Rev. Sociology* 34: 181–209. doi:10.1146/annurev.soc.33.040406.131740.

Paul, Laurie A. and Kieran Healy. 2018. "Transformative Treatments." *Noûs* 52 (2): 320–35. doi:10.1111/nous.12180.

Pearl, Judea. 2010. "An Introduction to Causal Inference." *International J. of Biostatistics* 6 (2): 7. doi:10.2202/1557-4679.1203.

Pedulla, David. 2014. "The Positive Consequences of Negative Stereotypes: Race, Sexual Orientation, and the Job Application Process." *Social Psychology Quarterly* 77 (1): 75–94. doi:10.1177/0190272513506229.

Porter, Theodore. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Quillian, Lincoln, Devah Pager, Ole Hexel and Arnfinn H. Midtbøen. 2017. "Meta-analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time." Proceedings of the National Academy of Sciences of the United States of America 114: 10870–75.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *J. of Educational Psychology* 66 (5): 688–701. doi:10.1037/h0037350.

Ho Daniel, and Rubin, Donald B. 2011. "Credible causal inference for empirical legal studies." *Annual Review of Law and Social Science* 7: 17-40.

Schaffer, Johnathan. 2005. "Contrastive Causation". *Philosophical Rev.* 114 (3): 327–58. doi:10.1215/00318108-114-3-327.

Simonsohn, Uri. 2016. "Greg Vs. Jamal: Why Didn't Bertrand and Mullainathan (2004) Replicate?" Data Colada, https://datacolada.org/51(accessed November 19, 2023).

Starr, Sonja B. 2016. "Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police." *University of Chicago Legal Forum* 2016 485–531.

Wodak, Daniel. 2022. "Of Witches and White Folks." *Philosophy & Phenomenological Research* 104 (3): 587–605. doi:10.1111/phpr.12799.

Starr Sonja 2024 The Magnet-School Wars and the Future of Colorblindness *Stanford Law Review* 76 1 161-268

Crabtree C, Kim J Yeon, Gaddis S Michael, Holbein J B, Guage C and Marx W W. (2023). Validated names for experimental studies on race and ethnicity. Sci Data, 10(1), 10.1038/s41597-023-01947-0

Kohler-Hausmann Issa 2018 Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination *Northwestern Law Review* 113 5 1163

Hu Lily, and Kohler-Hausmann Issa 2020 What's Sex Got To Do With Fair Machine Learning *ACM Conference on Fairness, Accountability, and Transparency* 2020 1

Kohler-Hausmann Issa 2024 What Did SFFA Ban? Acting on the Basis of Race and Treating People As Equals *Arizona Law Review* 66 305-356

Pager Devah, and Shepherd Hana 2008 The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets *Annual Review of Sociology* 34 1 181-209

https://doi.org/10.1017/lsr.2024.36 Published online by Cambridge University Press

Cases Cited

Brief for Economists as Amicus Curiae in Support of Petitioners, *Students for Fair Admissions v. President & Fellows of Harvard Coll.*, 600 U.S. 181, 289 (2023).
Comcast Corp. v. Nat'l Ass'n Afr. Am.-Owned Media, 140 S. Ct. 1009 (2020).
Students for Fair Admissions v. President and Fellows of Harvard College, 600 U.S. 181 (2023)
United States v. Duque-Nava, 315 F. Supp. 2d 1144, 1153 (D. Kan. 2004)
United States v. Johnson, 122 F. Supp. 3d 272, 331 (M.D.N.C. 2015).
United States v. Jones, 36 F. Supp. 2d 304, 307 (E.D. Va. 1999).
United States v. Payne, No. 12 CR 854 (CRN), slip. op. at 3 (N.D. Ill. Jan. 20, 2015)
Expert Report of Peter S. Arcidiacono at 17–21, *Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll.*, 397 F. Supp. 3d 126 (D. Mass. 2019) (No. 14-cv-14,176-ADB)
Floyd v. City of New York, 959 F. Supp. 2d 540 (2013)
Village of Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252
Washington v. Davis, 426 U.S. 229 (1976)

Issa Kohler-Hausmann is a Professor of Law at Yale University.

Lily Hu is an Assistant Professor of Philosophy at Yale University.

Cite this article: Hu, Lily and Issa Kohler-Hausmann. 2024. "What Is Perceived When Race Is Perceived and Why It Matters for Causal Inference and Discrimination Studies." *Law & Society Review* 1–26. https://doi.org/10.1017/lsr.2024.36