Signal Manipulation and the Causal Status of Race

Naftali Weinberger

October 2021

Abstract

Discussions of the causal status of race focus on the question of whether race itself can be experimentally manipulated. Yet many experiments testing for racial discrimination do not manipulate race, but rather a signal by which race influences an outcome. Such signal manipulations are easily formalized, though contexts of discrimination introduce an additional issue. Whether a signal counts as a signal for race is not merely a causal question, but depends on sociological and normative issues regarding discrimination. The notion of signal manipulation enables one to take these issues into account while still using causal counterfactual tests to detect discrimination.

1 Introduction

Discrimination is, in part, a causal concept. The claim that an individual was discriminated against based on her race entails that her race made a difference to how she was treated. Yet there is currently a great deal of perplexity surrounding causal claims involving demographic variables such as race. The problem arises within the methodology of causal inference, which evaluates causal claims using counterfactuals concerning the results of "manipulating" a purported cause for an individual, while keeping her other properties fixed. But it is doubtful whether there are hypothetical manipulations that change only a person's race, while keeping all other factors fixed. This concern has been taken seriously by social scientists studying discrimination. The three dominant proposals addressing this have been 1) to claim that the relevant manipulation is not on race, but on the discriminator's perception of race (Greiner and Rubin, 2011), 2) to claim that race is a cluster concept and that the relevant manipulations are on its components (Sen and Wasow, 2016), and 3) to deny that counterfactual causal methods are suitable for modeling discrimination (Kohler-Hausmann, 2018).

None of these approaches can account for the way that discrimination is empirically tested, even for simple experimental designs. Consider Bertrand and Mullainathan (2004), who sent out resumes assigned either stereotypically black

or white names and measured the difference in callback rates for otherwise similar candidates. What is manipulated here is not the person's race, but the way that information about race is transmitted to the employer. The "perception of race" proposal captures this, but introduces unnecessary complications about what it means to manipulate perceptions, while avoiding necessary questions about why the actual manipulation (i.e. on the name) is suitable. The second proposal would treat the racial-soundingness of a name as a part of the cluster-concept of race. This misses the fact that the name functions not as a *part* of race, but as a proxy for it. The third proposal emphasizes that the experiment cannot explain why certain categories such as race merit specific legal protections against discrimination, which make sense only in the context of broader social and historical patterns of discrimination. This is correct, but does not entail that the experiment is therefore not suitable for detecting discrimination.

A great deal of confusion can be avoided by gaining a clearer understanding of the type of manipulation involved in the experiment. The core idea is simple: although the experiment manipulates the name on the resume, this manipulation is informative only because the name is treated as a reliable signal for a person's race. This is in some sense obvious – the purpose of the experiment is not to test for *name* discrimination, but *race* discrimination. But this feature of the experiment has not been discussed systematically, and is in fact non-trivial to reconcile with features of causal methodology. Notably, manipulations, when formally modeled as interventions on variables in a causal model, make a variable no longer depend on its causes other than the intervention. However, the value of name on the resume derives from its carrying information about race, and this informational link would be destroyed by such an intervention. We see that such manipulations, which I call *signal manipulations*, behave differently from standard manipulations, and thus call for their own analysis.

One might worry that the signal manipulation proposal commits one to a problematically realist view of race. Assuming race is socially constructed, how can one talk about race independently of its signals? I will argue that this worry is ill founded: the social constructedness of race does not entail that the relevant construct must reduce to its various manifestations. What this means for the causal status of race is a complicated question that I will only begin to evaluate here. I will suggest that whether it makes sense to causally model race varies based on the scenario and the particular normative and/or empirical context. My proposal will help explain why one can talk sensibly about race causally in some cases and why it is problematic in others. This will offer a way towards a deeper exploration what makes race causally complicated, without throwing into doubt the ability to test for (at least certain kinds of) discrimination using well-designed experiments.

2 Manipulability

The starting point of this discussion is a rather technical methodological issue: Although it is common to test and analyze causal claims in terms of whether

one can influence the effect by manipulating the cause in a particular way, it is unclear if demographic variables such as race can be so manipulated (Holland, 1986; Glymour, 2007; Holland, 2008; VanderWeele and Robinson, 2014; Glymour and Glymour, 2014; Pearl, 2018). For readers interested in exploring race in all of its sociological and normative complexity, this technical focus might initially come as a disappointment. Nevertheless, there is a great deal to be learned from juxtaposing methodological questions about how to empirically detect racial discrimination with sociological/normative questions about what race is and what makes racial discrimination distinctively wrong. Two of the analyses we will consider (Sen and Wasow, 2016; Kohler-Hausmann, 2018) claim that manipulations on race make sense only if it is biologically real (as opposed to socially constructed). I believe that the connection between what race is and whether it is manipulable is not nearly as tight as suggested by these analyses. Clarifying these issues will require some background regarding what is at stake in debates over manipulability and why race is supposedly not manipulable. This section will provide this background.

I will use the terms "manipulation" and "intervention" interchangeably, though I note that the term "manipulation" often suggests a concrete change to a system, while the term "intervention" is preferable when discussing more abstract formal characterizations of such changes. *Ideal interventions* are a paradigm example of such a characterization. An ideal intervention on a variable sets that variable's value in such a way that it depends only on the intervention, rather than on its other causes. Such interventions are useful for addressing the problem of "confounding" (Pearl, 2009, ch. 6). As a standard illustration: a strong correlation between smoking and cancer might indicate not that smoking causes cancer, but merely that those with a propensity to develop lung cancer are also more likely to smoke. To rule out this possibility, one can intervene to make test participants smoke whether or not they would otherwise. In a *randomized control trial* this would be accomplished by randomly assigning subjects to either smoke or not smoke. The result of this is that learning whether a participant smokes provides one with no information about their properties – and thus about their propensity to develop lung cancer – and thus any remaining correlation between smoking and cancer reveals that smoking *causes* cancer.

Beyond their experimental utility, interventions have been central to the analysis of causation, since genuine causal relationships may be distinguished from merely probabilistic ones in terms of the former enabling one to predict the results of interventions (Woodward, 2003). Some philosophers might dismiss interventions as being of "merely" methodological significance, and thus of little relevance for understanding what causation is. Such philosophers would readily accept the existence of causes that cannot manipulated as evidence that the set of causes in the world is wider than the set of variables that one can intervene upon and thus might dismiss what follows as being of interest only to those adopting a controversial "interventionist" view of causation. As a preliminary response, it is crucial to clarify that the concept of an intervention applies even in cases where human agents cannot intervene for practical or ethical rea-

sons. Although the ethics research board won't permit interventions to make people smoke, it is clear enough what such an experiment would look like. The Federal Reserve would not randomly intervene upon the economy, but clever researchers will seek naturally occurring "instrumental" variables that play the same evidential role as interventions (Angrist et al., 1996). What matters in principle is how variables *would* respond to interventions, rather than whether such interventions are feasible.

Still, why think that the set of causes should be at all delimited by the set of manipulable variables? The position I adopt here is not that there cannot be non-manipulable causes (I remain agnostic on this point), but rather that certain types of non-manipulability threaten the intelligibility of viewing the variable in question as a cause. Taking a step back, questions about whether some relationship is causal only make sense relative to an implied contrast regarding what it would be for it *not* to be causal. In the well-understood case of the randomized trial, we are interested in whether someone's taking a drug brought about a certain outcome, or, instead, whether learning that someone took a drug is merely evidence that they would be likely to recover. The key to randomization is that because the treatment and control groups – e.g. the groups that received the drug as opposed to the placebo – differ only in their treatment status, one can take the outcome in the control group as reflecting the outcome that the members of the treatment group would have had (on average) had they received the control (and vice versa). This works only because it is possible to vary one property, while keeping the distribution of all other properties the same across the groups, and this possibility underlies interventions more generally. At the extreme, imagine that it were impossible to vary whether an individual was in one group or another without varying all properties of the individuals across the groups. It would arguably no longer make sense to ask whether the group an individual is in is causally or evidentially relevant to a particular outcome. Had she been in the other group, she would have been an entirely different individual. In such a case, the impossibility of characterizing an intervention on group status is not just a pragmatic limitation, but rather reflects that a standard basis for conceptually differentiating causal from non-causal relationships is inapplicable.

Discussions of whether demographic variables are manipulable are sometimes posed as metaphysical questions about whether it is possible to change someone's race or gender compatibly with their being the same person. But the real issue is not metaphysical identity, but rather whether the variable in question can be manipulated independently of other variables characterizing an individual (Weinberger, 2015). If it made sense to talk about changing a person's race independent of their other properties, then the question of whether they would count as the same individual would be of no further relevance to whether race is manipulable. Of course, it remains far from clear that it does make sense to talk about changing race in this way. This is why the question arises as to whether and how one can experimentally test for its purported effects. Nevertheless, the discussion has already yielded an insight that will be crucial for evaluating proposals for causally understanding race. Namely, whether it makes sense to

reason about race causally boils down to whether it is possible to reason about variation in race independently of variation in an individual's other properties.

In the context of empirically modeling racial discrimination there is often little need to answer unrestricted questions about how a person would have been different had they been of a different race. It would already be progress if we could clarify what it would mean for race to influence how a person is treated in particular discriminatory contexts. Even given a simplistic view of race, however, it is non-trivial to clarify how the effects of race are measured in a particular context. Marcellesi (2013) imagines a fictional experiment in which one intervenes on a fetus' genes in utero to change the future child's race (understood biologically). Even supposing that such an intervention made sense, it is clearly not the intervention of interest in typical contexts of discrimination. If one wants to know whether a job candidate who was denied a position last week was discriminated against, what we are manifestly *not* asking is whether she would have gotten the position had she spent her whole life as a person of a different race. Plausibly, such an intervention would change many facts about her life trajectory, making it unlikely that she would even end up applying for the job. We see that even if one is willing to conceptualize race in a limited way such that it can be manipulated, one needs to further show that the envisioned manipulation is relevant for establishing a certain type of discrimination.

In many contexts where a variable is said to be "non-manipulable" the issue is not that there is no intervention one could perform on it, but rather that there are too many distinct interventions. This might not be an issue when the intervened-upon variable can be characterized such in a way that its influence on its effects is invariant across different versions of the intervention. But many variables are not like this (Spirtes and Scheines, 2004). Regarding race, one can imagine a range of experiments testing the effects of race in a particular context, including interventions changing an individual's external appearance, interventions changing how that individual is viewed by others in a context, and perhaps even interventions on whether the individual lives in a society with racial disparities – to name just a few possibilities. Each of these corresponds to a different experiment with different effects, and consequently one cannot talk about intervening on race without clarifying which particular intervention one has in mind. This problem suggests its own solution (Woodward, 2003, p. 116-117): experimenters should limit themselves to only talking about race in the context of particular interventions on race, and to avoid discussing race without specifying particular interventions.

Although this disambiguation strategy is useful for generating manipulable variables, it should be viewed only as the first step towards understanding race and racial discrimination causally. Beyond specifying a manipulable race-related variable, one must defend that the relevant intervention is a suitable one for testing for racial discrimination. This is non-trivial, as we will presently see when we look at existing proposals in the literature for modeling (or rejecting) the manipulability of race. One basis for evaluating these proposals will be in terms of their ability to account for *audit studies* in which the experimenter varies a cue related to a person's race to see how doing so impacts whether they

are hired. Notably, Bertrand and Mullainathan (2004) sent out resumes randomly assigned names that are perceived as stereotypically white or black to see whether this impacted the fictional applicant's chance of receiving a callback (it did). Although this experiment only considers a particular type of employment discrimination in a limited context – e.g. one in which there is no in-person interaction – I take it to be a fairly clean cut test of racial discrimination. One would therefore hope that existing proposals for the manipulability of race could easily account for the relevant manipulation.

While the proposals to be discussed concern race, the issue of non-manipulability arises also for other demographic variables, such as gender and age (Fosse and Winship, 2019). Here I will focus on race, though in later discussions of socially constructed variables, an occasional example involving gender will help. Moreover, even with respect to race I will only consider certain types of purported effects. Specifically, I'll emphasize types of discrimination that involve agents or institutions "detecting" the races of individuals. This is what is at issue in many standard queries about employment discrimination and discrimination by police. We will see that not all discrimination is like this. Some discrimination (e.g. "redlining")[1] relies on indirect proxies for race (Prince and Schwarcz, 2019), and broader patterns of discrimination do not simply reduce to the actions of agents or institutions. These issues are pressing, and are more complicated that those considered here. This paper is motivated by my view that existing proposals cannot even account for simple tests of discrimination and that a better one is a prerequisite for clarifying more complicated scenarios.

3 Problems with Existing Solutions

3.1 Greiner and Rubin: Race as Perception of Race

Greiner and Rubin (2011) propose that instead of treating race itself as a cause, experiments on discrimination should be understood as manipulating the discriminator's perception of race. This proposal clearly gets something right. In audit experiments, one is interested in whether the candidate would have been called back had the employer *thought* they were of a different race, not whether they had *been* a different race. And it is a virtue of the proposal that it locates discrimination not in some allegedly intrinsic feature of the individual, but rather in how others respond to that individual. The problems with the proposal arise once one takes seriously the claim that what is being manipulated is a particular individual's perception. What does it mean to manipulate a mental state? What if there are multiple perceivers making the hiring decision? When is "the moment the decider first perceives a unit's race" (Greiner and Rubin, 2011, p. 775)? The point is not that one could not find answers to these nitty-gritty questions, but rather that these questions are irrelevant. In a

[1] "Redlining" refers to financial institutions treating zip codes predominantly containing minorities differently from zip codes that do not, thereby discriminating against minorities without explicitly taking minority-status into account at the decision-making level.

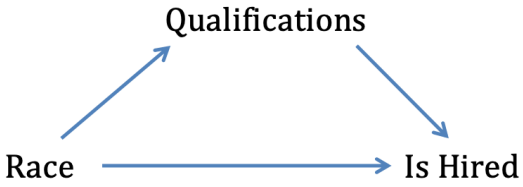

Figure 1: Discrimination DAG

well-designed audit study measuring callbacks, we simply do not need to know how many people were on the hiring committee to judge whether discrimination is occurring. The manipulation was not directly on anyone's mental state, but (e.g.) on the name on the resume.

Further questions arise if one takes the proposal as offering the variable *perception of race* as a replacement for *race*. Consider the plausible if simplistic model for audit studies in figure 1. The model posits a person's race as potentially influencing whether a candidate is hired via two paths (sets of connected causal arrows). First, prior discrimination may have led the candidate to be denied opportunities to develop job-relevant qualifications, resulting in the candidate being denied the job due to lacking these qualifications. This is represented by the indirect path from *race* to *qualifications* to *is hired*. Second, the candidate may be not hired due to the employer identifying their race. This is represented by a direct path in the model, but we could just as easily have included an indirect path through a variable for the name on the application. Pearl (2001) presents this model to illustrate that when evaluating discrimination, what matters is not whether race has *any* effect on being hired, but rather whether it influences being hired by the direct path (i.e. holding qualifications fixed). Note that in this model, one cannot just replace the variable *race* with *perception of race*. Although perception of race can plausibly describe what is happening along the direct path, the current hirer's perception of the candidate's race has nothing to do with why the candidate has her current qualifications. Although the issue here traces in part to the simplicity of the model, the lesson generalizes. One cannot simply replace race with some particular operationalization of race and assume that one's models will still be coherent. Even in relatively simple experiments, it will be important to keep track of both race as well as the way that it influences a particular outcome.

It would be too much to demand that in order to include race as a variable in a causal model, one must ensure that the defined variable reflects the concept of race in all of its sociological complexity. We should be willing to consider the effects of race at particular times, for particular durations, and in particular contexts, while acknowledging that any particular experiment will only correspond to one piece of the puzzle. One might imagine that doing so would make the methodologist's job very easy: for each experiment, one should just be clear how one is defining the race variable(s), perhaps with the appropriate subscripts. We

now see that this is not so trivial. There is obviously something correct in the claim that audit studies concern the perception of race. Nevertheless, Greiner and Rubin's proposal, as it stands, does little to spell out the details about how race is modeled within experiments. While models – causal or otherwise – run the risk of oversimplifying a complex scenario, the rewards come from their forcing the researcher to spell out their assumptions in a rigorous way. Greiner and Rubin fall short in this respect.

3.2 Sen and Wasow: Race as a Bundle of Sticks

Sen and Wasow (2016) propose that instead of attempting to define a single race variable that can be manipulated, race should instead be defined as a cluster property consisting of an aggregate set of elements, some of which can be manipulated in particular experiments. The guiding analogy is that race is a "bundle of sticks", where race corresponds to the whole bundle and the sticks correspond to its aggregate elements. As is evident from their choice of metaphor, their account is not designed to say anything about the relationship between the sticks and the bundle, or to provide a basis for theorizing which elements should count as being in the bundle at all. They do suggest that some elements of the bundle are more mutable than others, and thus more manipulable (507). Yet clearly the aim of the account is not to provide a unified framework for thinking about the relationship between the different elements of race, but rather to emphasize that race consists of many elements and that only a subset of elements will be under consideration in any particular study.

Sen and Wasow present their account as clarifying how the effects of race are tested in a range of experiments, including the audit experiments emphasized here. They go beyond Greiner and Rubin by distinguishing between race and its elements, but are unable to provide a satisfactory account of how these relate to one another. For instance, they explain audit experiments by saying that the manipulated "stick" (the cue) is a proxy for race. But proxies are not *parts* of the entities they represent, but distinct from them. Such mixing of metaphors would be harmless if Sen and Wasow offered an independent account of what determines which sticks counts as part of the bundle, but they do not. This leaves them without resources for explaining why the manipulations performed in discrimination experiments are the relevant ones to perform. The actual manipulations are on the sticks themselves, while the bundle itself is an idle wheel doing no explanatory or modeling work. In contrast, talk of proxies correctly captures the fact that audit studies vary cues such as the name on the resume to alter the racially-relevant information that is transmitted to the hirer. This, of course, invites the question of what it means for a cue to carry such information. The proxy metaphor nevertheless contrasts favorably with that of a bundle of sticks, which by design says nothing about the sticks' relationships to each other or to the bundle as a whole.

For other types of experiments, Sen and Wasow describe the sticks not as proxies for but as constitutive elements of race. But the shortcomings are similar: the proposal does not allow one to ask what makes something a constituent of race. One might suppose that this limitation could be rectified by appeal to the philosophical literature presenting constitutive explanations – in which components jointly explain the functioning of a whole mechanism – as an supplement to standard causal explanations (Craver, 2007). In the event that this paragraph is not cut as part of the reviewing process, I feel compelled to share my view that this literature is a muddled mess that pretends to be rigorous, while failing to solve any problem that is not of its own making.[2] While talk of constitutive explanations is now too entrenched within philosophy of biology and neuroscience to disappear any time soon, I would seek to dissuade methodologists who are unsatisfied with causal treatments of race from presuming that this literature provides novel tools for addressing their problems.

I suspect that the appeal of Sen and Wasow's position derives from an underlying assumption that since race is a social construct, one cannot legitimately model it as independent of its social manifestations. My suggestion that the manipulated variables are not parts of race but rather proxies would thus be rejected as relying on a contentious claim that race has an existence independent of its proxies. In section 5, however, I will argue that this underlying assumption is wrong. The social constructedness of race does not entail that race should not be modeled independently of its proxies.

3.3 Kohler-Hausmann: Abandon the Causal Approach

Sen and Wasow take for granted that whether race can be manipulated is closely related to whether race is a biologically essential feature of individuals, rather than socially constructed. Kohler-Hausmann (2018) provides an argument in favor of there being such a connection. Specifically, she argues that counterfactual causal tests for discrimination are incompatible with understanding race as socially constructed. Her view overlaps with Sen and Wasow's insofar as she talks about how race is "constituted". But her aim in doing so is not to decompose race into components that can be individually manipulated, but rather to argue that because it is constituted by a complex set of social relationships, localized counterfactual tests are unsuitable for detecting racial discrimination in experimental contexts. Kohler-Hausmann's discussion contains a wealth of insights that will be central to the subsequent discussion. In this section, however, my aim is to rebut her claim that the counterfactual approach presupposes the biological reality of race.

To illustrate her thesis that counterfactual approaches are inappropriate for evaluating socially constructed categories, Kohler-Hausmann describes a hypothetical society stratified into two groups: Royals and Non-Royals. Royals are identified by their wearing purple capes and carrying sticks, and are considered

[2]Romero (2015), while sympathetic to the mechanistic project, compellingly argues for the inconsistency of its most rigorous existing formulation. Weinberger (2019) highlights why such inconsistencies undermine the explanatory aims of the project.

so above Non-Royals that the latter are expected to step off the sidewalk when a Royal approaches. An anthropologist observing this society would miss the significance of the Royal/Non-Royal distinction if they myopically focused on the effects of capes on sidewalk behavior. The counterfactual "If I changed an isolated trait about a person (cape, stick) and nothing else changed about that person, would other pedestrians have remained on the sidewalk?" (p. 1180) locates the disparate treatment in the behavior of individuals, while it is better understood as a reflecting the social stratification of the society. Put differently, the answer to the question "why did the Non-Royal step into the road" is not, "because she encountered a cape-wearer", but "because she encountered a Royal" (cf. Dembroff et al., 2020). Being a Royal does not reduce to the features used to identify Royals, but is a fact about the status attributed to different groups within a society. Similarly, race is a social category, and Kohler-Hausmann sees the counterfactual approach as a misguided attempt to reduce it to superficial features such as skin color.

In modeling discrimination, the anthropologist's lens is not the only one that matters. Intervening to give a Non-Royal a cape and stick and seeing if other Non-Royals defer to them would not tell the researcher why Royals are privileged within the society. But it is a perfectly good test of whether Non-Royals are being discriminated against in the particular context. Such a test would not resolve the question of whether such discrimination is bad or what makes it so, and thus would be of limited use to a lawmaker deciding whether to design legal protections against it. But it would nevertheless be a vital test to conduct if one was uncertain whether Royals and Non-Royals were being treated differently on sidewalks. Contra Kohler-Hausmann, a category's being socially constructed does not mean that counterfactual tests are ill-suited for testing for discrimination based on that category.

Throughout, Kohler-Hausmann conflates the issues of whether causal counterfactuals provide an analysis of what discrimination is and why it is bad with the issue of whether such counterfactuals are relevant to *detecting* discrimination, which is what she claims to be addressing. Before elaborating upon this criticism, I should emphasize that there are cases in which counterfactual tests are used to address the former issues, and for those her objections hit their mark. Dembroff, Kohler-Hausmann, and Sugarman (2020) discuss a lawsuit pertaining not to race but sex, and in particular whether protections based on sex also protect individuals based on sexual orientation. In particular, if a male employee is fired for being married to a man, do we say either that A) he was discriminated against based on sex, because had he been a woman married to a man he would not have been fired or that B) he was not discriminated against based on sex, but on sexual orientation, since had he been a woman in a heterosexual marriage he would not have been fired. Each option appeals to a different counterfactual, and which counterfactual to choose cannot be resolved by appeal to value-free facts about the individual, but rather to social questions about whether sexual orientation is part of the social category of sex, as well as normative questions about why sex merits legal protections against discrimination. Moreover, social context is necessary not only for selecting protected categories, but also for eval-

uating what counts as part of a category. A study that sent male and female candidates to a job interview wearing the same dress (Kohler-Hausmann, 2018, p. 1216) would not be a suitable test for sex discrimination, as the candidates would differ not just in their sex but in their conforming to sex-based norms. Seeing this requires knowledge of the role of sex within society.

Kohler-Hausmann emphasizes that the counterfactual approach sheds little light on what race is or what makes discrimination based on race – as opposed to, e.g., having freckles – sufficiently bad to merit special constitutional protections. The answer to these questions is found not in facts about individuals, but requires reference to broader social and historical patterns of discrimination. Nevertheless, it does not follow that *given* assumptions about which categories are protected, one cannot use counterfactual tests for detecting discrimination. And it is not clear how one could do without such tests. An individual's claim to have been racially discriminated against would be undermined if they would have been treated similarly even had they been of a different race.

4 Signal Manipulation

In testing whether Non-Royals defer to Royals, one does not manipulate whether someone is a Royal, but whether someone has a cape and stick. This test only works, however, if the cape and stick are reliable signals for royalty. If the Non-Royal were not fooled into identifying the caped individual as Royal, she would not defer to her. That said, if the disguise *is* convincing, the test does establish an effect of royalty in this localized context. This effect is just one node in the nexus by which the society reinforces its social structure. Yet even this localized example has structure that eludes existing approaches to manipulability. The test described requires that one attends to both the Royal status and its signal, as well as their systematic relationship. Greiner and Rubin focus only on the signal. Sen and Wasow cannot account for the systematic relationship. Kohler-Hausmann recognizes the complexity and non-manipulability of status, but denies that the effects of status are tested via the manipulation of signals.

The robe and stick are reliable signals for royalty insofar as they counterfactually track one's royalty status. Outside of the experiment, they correspond to what one would have if and only if one were a Royal. Similarly, in Bertrand and Mullainathan (2004), the name on the resume is only significant insofar as it reliably tracks the applicant's race. The study involves holding the individual's qualifications and non-racially-identifying factors fixed while changing the signal by which information about their race is transmitted.

Schematically, we can describe a signal-based manipulation with three variables, C for the category (e.g. race), S for the signal, and Y for the outcome (i.e. the effect). S reliably tracks C, and this relationship obviously calls for further elaboration. From the modeling perspective, the most crucial feature of this relationship is that one can specify how S would counterfactually vary given different values of C. A signal-based manipulation involves manipulating S. But unlike with ideal interventions, one does not do so to render it fully in-

dependent of its prior causes. Rather, one manipulates it to make it vary in the way it would were C to have one value as opposed to another. So if C has two values, c and c', which are reliably tracked by the values s and s', respectively, then signal-based manipulations for the influence of C on Y would correspond to the result of manipulating S from s to s'.

In Bertrand and Mullainathan (2004), C is race (black or white), S is the name, where certain names are treated as reliable signals of race, and Y is whether the applicant gets a callback. To test for whether a black candidate was discriminated against for being white, one changes the name from that which the candidate would (and does) have as a result of being black to a name they plausibly would have were they white. If this change in the name leads to a change in the probability of getting a callback, this is taken to show that *race* influences callback. In this manner, one tests for the effect of C on Y by manipulating S.

For the rest of the section, I will further explore this proposal by pretending that C were an ordinary cause of S and that this accounted for the counterfactual dependence of the latter on the former. This resulting causal model contains the path $C \rightarrow S \rightarrow Y$. In such a model, one could determine the effect of C on Y via a signal manipulation of S. The ability to do so is at the center of *causal mediation techniques* (Pearl, 2001; Imai et al., 2010), which measure the way that an effect variable would respond to a change in a cause variable, were that change to be transmitted via some causal paths but not others (Weinberger, 2019). For models with multiple paths between the cause and effect, the definitions for the various effects along the different paths (e.g. direct and indirect effects) can become somewhat involved. For our purposes, the key facts are that A) in the case where the *only* path between C and Y is via S, the total effect of C on Y will just be its indirect effect via S, and B) indirect effects are measured by signal manipulations of the mediators. This suffices to show that the effect of C on Y can be established via a signal manipulation.

Recall that ideal interventions disrupt the relationship between the intervened-upon variable and its causes other than the intervention. If one manipulated the signal in this way, it would no longer function as a signal. Signals convey information – typically about their causes or effects of their causes – and ideal interventions break the causal connections by which this information is transmitted. An intervention that assigns a person a name independent of any other fact about them will, by design, render the name informationally independent of that person's race. Signal manipulations do not work like this. In selecting the contrastive values to which one intervenes to set the signal, one assigns the values that would obtain given different instantiations of category C. In this sense, the signal behaves as if it were tracking C. Nevertheless, it is not tracking C (within the experiment), since its variation is not due to variation in C. Although one changes the name on the resume from a black-sounding to a white-sounding name, the applicant remains black.

Thinking about certain tests of discrimination as testing the "effect" of race on some outcome via manipulating racial signals has several virtues. Most obviously, it captures the fact that the manipulation is not on race itself, but

rather on a variable by which race influences an outcome. Nevertheless, the effect of interest is not that of the manipulated variable, but rather of race. This is reflected in the modeling insofar as one manipulates the signal in a way as to mimic counterfactual changes in the value of C. More subtly, the fact that the signal only matters as a proxy for the category is reflected in the fact that the same effect could in principle be measured by manipulating distinct signals. Consider again the model from figure 1, in which race influences being hired via qualifications and via a direct path. Assuming that whether the employer discriminated depends on whether there is an influence via the direct path, it does not matter precisely which signal is manipulated, provided that one manipulates a set of variables that are sufficient to account for the influence via the path (i.e. if one were to include those variables in the model, there would be no remaining direct path). This reflects the idea that it simply shouldn't matter precisely when a particular hirer perceived the applicant's race or whether there was one or multiple hirers. What matters is whether race made a difference for employment not going through qualifications, not the precise avenue of influence.

Even in experiments in which the signal is is not directly manipulated, the concept of a signal manipulation may be relevant. Consider the experimental design of Grogger and Ridgeway (2006), which compared the racial disparity in police traffic stops during day and during the night. The underlying assumption was that during the night it would not be possible for police to identify the driver's race prior to stopping them, and thus the nighttime stop rates could be used as a benchmark for what the disparity would be in the absence of discrimination. If the disparity were higher during the day, this would be evidence that race was being taken into account in decisions about which cars to stop. Here there is no manipulation to make drivers appear to be of a different race. One would be hard pressed to localize the variable being manipulated. But the case is easily accounted for in terms of signal manipulation. If the assumption underlying the design is correct, then the signals about drivers' race are transmitted during the day, but not at night. The experiment works by manipulating whether the signal is transmitted.

We see that modeling the interventions as signal manipulations seems to provide a fruitful way for thinking about how audit studies test for racial discrimination. Yet the relationship between race and its signals has yet to be properly explicated. Talk of the signal "reliably tracking" the category has a causal ring to it. But without specifying what the race variable designates we cannot say if it is manipulable, and we have thus replaced the question of whether race is a cause of an outcome with that of whether race is a cause of its signals. This might not initially seem like progress. The reason that it is is that there is a story to be told about why certain variables are only manipulable via their signals. Telling this story will bring into focus the genuine reasons why interventions on race are problematic.

5 Race and Manipulability: A New Lens

Earlier I noted a common – if implicit – assumption that since race is just a social construct, causal descriptions of it must reduce it to its social manifestations. This is what makes it sound plausible to say that race *just is* perception of race, or that it is nothing more than a bundle of sticks. But the underlying assumption is false. That race is a social construct does not imply that it reduces to its social manifestations. Seeing why is crucial for understanding the ways in which it does and does not behave like a cause.

Consider the question of whether a work of art was bought for a high price because it was authentic (as opposed to a forgery). The fact that original paintings are considered so much more valuable than near-indistinguishable forgeries is a societal fact about how art is valued. In this sense, the authenticity of a work of art is a social construct. There is little doubt about whether authenticity influences price and there are strong institutional mechanisms for ensuring the reliability of signals about a work's authenticity. Yet, if there were any doubt about the effects of authenticity on price, it would be impossible to test this by intervening to transform a forgery into the real thing. Rather one would have to manipulate the signals for authenticity and see the effect on price. But this does not mean that authenticity reduces to these signals. Although the value of authenticity is a social construct, *within* the socially-constructed theory of value what matters ultimately is not whether people believe an artwork is authentic, but whether it *is* so.

Something subtle is going on here. On one level, it makes no sense to talk about the value of authenticity independent of the beliefs of the community of individuals who value it. Yet I claim that what matters is whether a work is authentic, rather than merely whether individuals believe it to be so. To see why this is not a contradiction, one must distinguish between whether the concept of authenticity depends on the existence of certain beliefs and the *content* of those beliefs. Part of this content is that there is a fact of the matter about whether a work of art is authentic, whether or not people in fact believe it to be so. Accordingly, it is intelligible to talk about people having false beliefs about a work's authenticity. Understood properly, this way of talking allows for a distinction between authenticity and beliefs about authenticity, without committing one to the view that authenticity exists independently of the society within which it is valued.

The general point is that the claim that a property is a social construct whose influence can only be tested by manipulating its signals does not entail that it should be understood as being equivalent to those signals. One can (and should) distinguish between the property itself and the signals by which it is detected.

With this in mind, let's revisit the suggestion that race should be understood as perception of race. This suggestion is potentially ambiguous. On the one hand, it might just indicate that race is not an intrinsic feature of an individual, but depends on how one is perceived within a society. Alternatively, it might suggest that all that matters is whether a person is perceived as being of

a certain race, whether or not they are of that race. This second interpretation is much less plausible, and one should not take it as following from the first. If a white candidate were denied a job because the employer mistakenly believed they were black, we would not say that they were discriminated against based on race. This is so even though their perceived race might match that of a black candidate who was discriminated against based on an accurate perception of their race. This feature of linguistic practice makes sense within a society in which people's perceptions of race fairly reliably track the way that an individual is categorized within the society, and where the ability to detect race in this manner is an important component of how discrimination functions within that society. The point is not that individuals are never mistreated based on false racial identifications, but rather that these aberrant cases are not really what is at issue when studying (or designing legal protections to counteract) the phenomenon of people who are discriminated against based on accurate perceptions of their socially-constructed race.[3]

There is an apparent tension between my claim that signals must reliably track the category and the fact that signal manipulations vary the signal without also changing the category. In the artwork example, when one tests the effect of authenticity by seeing what buyers would offer for an authentic painting that is listed as a forgery, one is making it the case that the signal *fails* to reliably track what it is supposed to. The key to resolving this tension is that the possibility of altering the reliability of the signal within the contrived scenario of the experiment does not undermine its reliability outside of the experimental context. In fact, the reason why the signal manipulation is deemed relevant to understanding the behavior of the category outside of the experiment is precisely because one manipulates the signal in the experiment to what it would be in the real world under the condition in which the category would be different from what it is. The experiment in which a genuine artwork is signaled to be a forgery is informative precisely because the signal in the experiment corresponds to the one that would be manifest in the actual world were the artwork to be a forgery.

Of course, the question of what makes someone count as being of a certain race is by no means as clear cut as that of what makes a work of art authentic,

[3]The idea that socially-mediated causal relations implicitly rule out certain types of aberrant cases is evident from an example by Prescott-Couch (2017). Georgina is a student who is stigmatized as a result of having a birthmark covering half her face and as a result of the stigmatization is distressed. Intuitively, the birthmark causes her being distressed via – and, we could imagine, *only* via – stigmatization. So it seems like the causal model should be *birthmark* $\rightarrow$ *stigmatization* $\rightarrow$ *distress*. This model entails that if one were to hold stigmatization fixed, intervening on the birthmark would not have any effect on distress. Prescott-Couch then considers the scenario in which one intervenes to remove Georgina's birthmark, while holding fixed the fact that she is mercilessless teased by her classmates for having one. The Twilight Zone-esque scenario in which she is teased for a birthmark she doesn't have would be distressing indeed – and presumably in a different way and to a different degree than in the scenario where she in fact has a birthmark. Nevertheless, Prescott-Couch compellingly resists the conclusion that (contra the proposed model) the birthmark *does* have a direct effect on distress not via stigmatization. Rather, to the extent we are interested in the social effects of having a birthmark in the ordinary scenario, the aberrant Twilight Zone scenario need not be considered.

and we should expect a fair amount of vagueness in racial characterizations. But none of this undermines the possibility of distinguishing between what race is (within a society) and how it is perceived in particular contexts. Race corresponds to a status that individuals are understood to have independent of its particular manifestations by which it is detected. The racial signals have the effects they do *because* they are taken as signals for the relevant social status. Even though one can only manipulate race via its signals, this does not mean that race is nothing over and above its signals.

Kohler-Hausmann presents a criticism of the Greiner and Rubin proposal that may seem to carry over to the signal manipulation approach. She claims that if race is a multifaceted social construct, then the perception of race will be so as well, implying that their proposal only gets off the ground if one identifies race with superficial features such as skin color. Whatever the merits of this criticism for the Greiner and Rubin proposal, it does not touch the signal manipulation approach. Even granting that race is a social construct, one is not required see race as identical to – or constituted by – the signals by which it is identified.

6 When is it Fruitful to Model Race Causally?

Contra Kohler-Hausmann, the fact that race is a social construct does not rule out the possibility of detecting race by intervening on its signals. In fact, her Royals example provides a good illustration of precisely such an intervention. We've seen that the relationship between race and its signals does not behave like an ordinary causal relationship, but I've argued that this should not stand in the way of using signal manipulations to test claims about the effects of race. In this section, I consider one final objection, which claims that the challenges that stand in the way of manipulating race carry over to manipulations on its purported signals. Through addressing this objection, I will provide a clearer characterization of when it is and is not fruitful to model race causally.

Consider Kohler-Hausmann's example of a test of gender discrimination in which male and female candidates come to an interview in the same dress. This test fails because in comparing genders we need not merely vary the gender of the applicant, but also the associated gender-conforming behaviors. This furnishes a nice example of how manipulations on gender may call for multiple simultaneous interventions going beyond gender in its most narrow sense. Moreover, the knowledge that one needs to manipulate the candidates' outfits comes from a broader understanding of how gender functions in society. This is worrisome for the signal manipulation proposal, since the question of which factors need to be varied to count as a suitable manipulation on race or gender applies just as much to their signals as to the categories themselves. Bertrand and Mullainathan's audit study wouldn't work if, after manipulating the name, the candidate's race could still be inferred from the neighborhood that they grew up in or the college they attended. While the signal manipulation proposal promises to create space between questions of what race is and questions of how one tests for its localized

discriminatory effects, perhaps this is just an illusion. Perhaps all of the issues that arise in trying to model manipulations on race itself simply reappear when one seeks to spell out all of the variables that must be manipulated as part of a signal manipulation.

This worry has some merit and in fact provides clues as to why in some cases it may no longer make sense to think about certain demographic variables causally. Fortunately, however, the problem does not generalize to all cases. Distinguishing between the cases in which the problem can and cannot be addressed will thus be helpful for seeing why certain demographic variables only sometimes can be treated causally.

Consider the question of whether in the first 2016 presidential debate Hillary Clinton was treated unfairly because she was a woman. Although Clinton was widely perceived as coming out ahead in the debate, Trump was deemed by many observers to have performed adequately. Given the obvious disparity between the knowledgeability of the candidates on issues of substance, it was fair to wonder if the debate would have been deemed to be as close as it had been had the gender roles been reversed. To test this, some academics commissioned a play in which each candidate was played by an actor of the opposite gender, but otherwise copied the candidates' words verbatim, along with the accompanying gestures (Reynolds, 2017). Surprisingly, many attendees of the play reacted more negatively to the male "Clinton" than they had to the original. Why was he smiling so much? And didn't he seem a bit effeminate? These reactions are informative, but we should be cautious in drawing inferences about the effects of the candidates' genders. How audience members react to a candidate's smiling, for example, is not independent of her gender.

In this particular example, there may be no meaningful way to resolve questions about the causal effects of Clinton's gender. Once it becomes clear that testing for the effects of gender would involve manipulations on any behaviors or gestures that would be interpreted in a gendered way, one begins to wonder what *wouldn't* have to be manipulated. As the list of signals to be manipulated becomes large and open-ended, the gap between the questions "how would she have been treated had she been of a different gender?" and "how are politicians of different genders treated within a society" narrows to the point of imperceptibility. To the extent one cannot distinguish between these two questions, the informal and plausible claim that Clinton's gender causally influenced how she was treated cannot be given a rigorous justification.

Such issues are in no way limited to gender. Imagine an audit study involving an in-person interview instead of just a resume submission. If the employer selects a white candidate over a black candidate on the basis that the former was more assertive, we should not immediately assume that race was not in play. Even if we could know that assertiveness was the only difference, the employer might have interpreted assertiveness more negatively when coming from a black candidate. The Bertrand and Mullainathan (2004) study is thus only able to work as well as it does given a contrived scenario in which one can really make it so that the employer has no knowledge of the candidate's race (which itself is randomly assigned to a resume). The lesson to draw from this is neither that

the study has no relevance to in-person scenario, nor that there are not sensible ways to further explore the causal impact of race in the in-person case. It seems plausible that there exists a spectrum of scenarios ranging from the cases where it is most feasible to isolate an effect of race via manipulation of a signal to those in which race operates in such a holistic manner that such manipulations are no longer coherent.

Earlier I suggested that the issue with manipulability of race wasn't metaphysical identity, but whether race could be manipulated independently of other properties of an individual. In the context of signal manipulation, one does not intervene on race, but rather on its signals. So the issue instead is whether one can manipulate some signals independently of others. The justification for manipulating some signals but not others need not be that the non-manipulated signals do not relate to race. In the simple hiring model in figure 1, the applicant's qualifications do convey information about her race. The reason for only considering the direct path in that scenario is that, in the particular context, we would likely not treat an employer who took these qualifications into account as discriminating against the candidate. In other words, although all paths in the model are related to race, only some are relevant for establishing discrimination in the particular context.

We see that whether it makes sense to evaluate a particular demographic variable causally depends not just on the variable, but also on the particular query. The more one asks specific questions about how particular individuals were treated in localized and well-defined contexts, the better the chances one will be able to provide a causal analysis. But there will, of course, be lots of borderline cases and any empirical analysis will rely on certain idealizations and simplifying assumptions about the influences of a variable in a context. In light of this, the most pressing question is: what do we gain by this modeling exercise of trying to model race causally? One could imagine Kohler-Hausmann granting a great deal of what I've said here, but taking it as showing that the counterfactual analysis only applies to idealized experimental scenarios. If what we really care about is the broad sociological concept of race, and I've granted that such sociological factors are necessary for specifying the manipulations required for testing discrimination, why go through the effort of seeking out an idealized case to which causal concepts apply?

My answer to this is that conceding that race cannot be modelled causally would be crippling for any attempt to design legal protections against racial discrimination. In typical hiring scenarios, there is nothing barring an employer from eliminating a candidate from consideration based on brute dislike. To designate race as a protected category, one must be capable of distinguishing cases in which candidates are denied a position *because* of their race, rather than as a result of such brute dislike. In the extreme scenario – the one that I've deemed non-causal on the basis that one cannot differentiate among the signals that are and are not signals for race – such a distinction is no longer tenable.

A remaining concern is that without being able to specify hypothetical manipulations of race, one cannot address concerns relevant to confounding. This is especially pressing in contexts where one wants to separate the influence of race

from closely related social categories such as class. Unfortunately, the signal manipulation proposal contains no means for answering this question. The way one chooses to manipulate a signal *presupposes* certain ways of thinking about the race/class distinction rather than providing a novel way to draw it. But it is not really clear that subtle issues regarding if and how one can differentiate the influences of race and class within a society are to be addressed with particular experiments, or even sets of experiments. Here the tool kits of sociologists and historians seem more directly relevant. I by no means want to deny the urgency of these questions, or to foreclose the possibility that experiments have some role to play in addressing them. Fortunately, however, we do not need to definitively resolve them in order to make progress in the design of experiments for detecting discrimination.

7 Conclusion

Instead of answering the question "is race a cause?", I have here sought to determine when it is fruitful to reason causally about race. Although many will still desire an answer to the first question, I see the type of project started in this paper as being essential for understanding both why manipulability matters (to the extent it does) as well as the specific problems arising when modeling race. Manipulability matters to the extent that when a variable cannot be manipulated independently of others, the distinction between causal and evidential relevance collapses. In the context of race, the question becomes one of whether it is possible to distinguish between someone's race being evidence for an outcome and its bringing about that outcome. It appears that, depending on the context and the query, this distinction is only sometimes coherent, but, to the extent it is, causal reasoning has a role to play. An account of the causal status of race should have something to say about the cases that can and cannot be modeled causally.

A general challenge in researching the causal status of race involves determining how to combine the methodological literature on causal modeling with social scientific work on race. Both Sen and Wasow (2016) and Kohler-Hausmann (2018) draw on conclusions about the relationship between the manipulability of race and its status as a social construct. For Sen and Wasow, the socially constructed nature of race is reflected in its being a cluster concept. For Kohler-Hausmann, the key fact is that race can only be fully understood as part of a socially constructed system of meanings. The forgoing discussion provides reason to be cautious of drawing any direct inferences from the social constructedness of race to its (non-)manipulability. *That* race functions as a social construct is less important than *how* it does. I've proposed that the reason that race eludes manipulability is that it is understood within society as having an existence independent of its particular manifestations. This is a particular claim about how the concept functions within the society and it is further important for understanding why it would be limiting to insist that one treat it as reducing to the signals that are in fact manipulated.

In the process of working through the philosophical and methodological issues involved in causally interpreting race, it is important that this research should aid rather than hinder experimental work on detecting racial discrimination. Audit studies offer a good experimental design for testing for a particular type of discrimination in particular contexts. The fact that the proposals reviewed in this paper make such studies more difficult to interpret suggests that something has gone wrong. The signal manipulation proposal provides a framework within which to continue to explore the thornier methodological questions without losing sight of what goes right in existing experiments. It straightforwardly yields the result that one can test for the effects of race by intervening upon its signals, but that one does not need to understand the effect as being simply that of the signals themselves. This is an important feature of audit studies that has been lost in prior theorizing, and the question of how much the signal manipulation proposal can be generalized beyond simple experimental designs is a fruitful starting point for further work. The category-signal relationship is not an ordinary causal one. But the conclusion to draw is not that we should avoid discussing such relationships, but rather that by analyzing them more carefully, we can gain insight into the particular challenges in causally understanding race.

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Dembroff, R., I. Kohler-Hausmann, and E. Sugarman (2020). What taylor swift and beyoncé teach us about sex and causes. *U. Pa. L. Rev. Online 169*, 1.

Fosse, E. and C. Winship (2019). Analyzing age-period-cohort data: A review and critique. *Annual Review of Sociology 45*, 467–492.

Glymour, C. (2007). Statistical jokes and social effects: intervention and invariance in social relations. In A. Gopnik and L. Schulz (Eds.), *Causal learning. Psychology, philosophy, and computation*, pp. 294–300. Oxford University Press.

Glymour, C. and M. R. Glymour (2014). Commentary: race and sex are causes. *Epidemiology 25*(4), 488–490.

Greiner, D. J. and D. B. Rubin (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics 93*(3), 775–785.

Grogger, J. and G. Ridgeway (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association 101*(475), 878–887.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

Holland, P. W. (2008). Causation and race. *White logic, white methods: Racism and methodology*, 93–109.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological methods 15*(4), 309.

Kohler-Hausmann, I. (2018). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev. 113*, 1163.

Marcellesi, A. (2013). Is race a cause? *Philosophy of Science 80*(5), 650–659.

Pearl, J. (2001). Direct and Indirect Effects. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge: Cambridge University Press.

Pearl, J. (2018). Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference 6*(2).

Prescott-Couch, A. (2017). Explanation and manipulation 1. *Noûs 51*(3), 484–520.

Prince, A. E. and D. Schwarcz (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev. 105*, 1257.

Reynolds, E. (2017). What if donald trump and hillary clinton had swapped genders. *NYU News (online)*.

Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese 192*(11), 3731–3755.

Sen, M. and O. Wasow (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science 19*, 499–522.

Spirtes, P. and R. Scheines (2004). Causal inference of ambiguous manipulations. *Philosophy of Science 71*(5), 833–845.

VanderWeele, T. J. and W. R. Robinson (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.) 25*(4), 473.

Weinberger, N. (2015). If intelligence is a cause, it is a within-subjects cause. *Theory & Psychology 25*(3), 346–361.

Weinberger, N. (2019). Mechanisms without mechanistic explanation. *Synthese 196*(6), 2323–2340.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford University Press.