Race Causality Discrimination – What Have We done?

Marcello Di Bello - ASU - Fall 2023 - Week #15

These are some scattered thoughts about what we have learned in this seminar, in part old thoughts drawn from earlier handouts and in part new thoughts.

Holland and Wooodward

We started out with a classic article by Paul Holland[1] that outlines the Rubin model of causality. This is a popular framework that describes how many statisticians thinks about causality. For Holland, a cause operates relative to another, competing cause. The effect of a cause on an individual unit is measured relative to a response variable of interest. The causal effect of t (relative to c) on unit u (measured by Y) is expressed by the difference:

$$Y_t(u) - Y_c(u)$$

In words, the causal effects of treatment t is the difference between the actual value of $Y(u)$, that is, the value of Y when unit u is exposed to the treatment t and the would-be value of $Y(u)$, that is, the value that Y would have taken had u been exposed to control c instead of treatment t. So, causal effects depend – at least conceptually – on a very specific counterfactual claim.[2]

For a different, yet similar account of causality, we turned to the philosophical literature, in particular, James Woodward's manipulability theory of causation.[3] The basic idea is this. For X to be a cause of Y, the following counterfactual should hold: if the value of X were to change *as a result of an intervention*, then the value of Y would change, as well. If there is no intervention such that changing the value of X would also change the value of Y, then X isn't a cause of Y. Crucially, this is not a merely counterfactual account of causality. The requirement for X to be a cause of Y is not simply that, if the value of X were to change, then the value of Y would change as well. Key is the notion of an intervention. An intervention on a variable X in order to change another variable Y should be exogenous to the existing relations that hold between X, Y and other variables.[4]

Can RACE be a cause?

Can RACE[5] be a cause? Holland thinks the answer should be negative.[6] The causal effect of $RACE = b$ (compared to another $RACE =$

[1] Holland (1986), Statistics and Causal Inference, *Journal of the American Statistical Society*, 81(396).

[2] This difference cannot be directly observed because the same unit u cannot be both subject to treatment and control (at the same time); see the fundamental problem of causal inference. Under certain conditions, it is more common to study the *average causal effect* of a treatment t across units in a population U, that is, the difference $E(Y_t - Y_c)$

[3] James Woodward (2003), *Making Things Happen: A Theory of Causal Explanation*, 2003, Oxford University Press. We'll focus on chapter 2.

[4] Woodward writes: "we may think of an intervention on X with respect to Y as an exogenous causal process that changes X in such a way and under conditions such that if any change occurs in Y, it occurs only in virtue of Y's relationship to X and not in any other way." (p. 47)

[5] This follows Holland's convention of referring to the variable 'race' simply as RACE.

[6] See Holland (2003), Race and Cause, *Research Report*, January 2003 RR-03-03, ETS Educational Testing Service.

w) on a unit u as measured by a response variable Y would be defined as follows:

$$Y_b(u) - Y_w(u)$$

In words, the causal effects of $RACE = b$ is the difference between the actual value of $Y(u)$, that is, the value of Y when unit u is Black and the would-be value of $Y(u)$, that is, the value that Y would have taken had u been White instead of Black. Holland thinks that the latter counterfactual statement makes no sense.

If RACE cannot be a causal variable, this raises the more general question of what can count as a causal variable. Holland's account views causality at the level of individual units u, such as people, households, plots of land. It is crucial that individual units can be *exposed to causes*. Anything individual units cannot be exposed to will not be regarded as a cause in any meaningful sense. So Holland's criterion for being a causal variable is that the variable could—at least in theory, though not necessarily in practice—be a treatment in an experiment. This criterion of causality makes many variables non-causal. Test scores, age and gender, as well as race, no longer count as causal variables.[7] Individuals cannot be exposed to race or gender, and it is unclear how they can be exposed to income. Instead, a cure or a school program are clearly things individuals can be exposed to and thus can be regarded as causes.

These consequences of Holland's account of causality seem overly restrictive, or so some have argued.[8] Consider, for example, a society in which two (otherwise homogeneous) racial groups, A and B, are paid differently by an employer. The assumption here is that the two groups are otherwise indistinguishable except for their race. A plausible explanation is that employers value the work of A's differently from the work of B's and decide to pay them differently. If A and B were races, race would play a causal role in this scenario. However, one might object that race itself does not cause differences in wages. Rather, what causes differences in wages is what employers believe about people's races. So, the causal mechanism would comprise *perceptions* of race, not race itself. Still, perceptions of race should arguably be caused by race and thus race would—ultimately—be the cause. If, on the other hand, perceptions of race are not caused by race, it is unclear how the employer's behavior can ever count as racial discrimination.[9]

[7] "From this point of view, attributes of individuals such as test scores, age, gender, and RACE are not causes and their measurement does not constitute a causal variable."(p. 9)

[8] Marcellesi (2013), Is Race a Cause?, *Philosophy of Science*, 80(5).

[9] Presumably, if a theory of causality rules out racial discrimination, it cannot be a good theory.

Audit studies and Hu's critique

Audit studies, also known as field studies, are often used to study racial and gender discrimination. Here is example of an audit study about racial discrimination in hiring.[10] A bank of resumes is created

[10] See Bertrand and Mullainathan (2004), Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination, *The American Economic Review*, 94(4), pp. 991-1013.

for different typologies of jobs. The bank contains several resumes divided into high- and low-quality. Resumes of the same quality should be judged by employers as equally good matches for different types of jobs. For each job ad, a pair of high-quality and a pair of low quality resumes from the bank—so four in total—are sent to the same employer. The only relevant difference between the pairs of resumes is the name: one resume is assigned a typical White name and the other is assigned a typical African-American name. The names are randomly assigned. Callback rates are then compared across resumes associated with White and Black applicants.

Suppose the callback rates are higher for White than Black applicants (as they are experimentally). Would this be evidence of racial discrimination? Since the names were randomized, the White and African-American resumes should be ranked the same on average. There could be alternative explanations, but overall this would seem compelling evidence of racial discrimination.

Audit studies mimic the randomized controlled trial in the social sciences where observational studies are more common. The assumption is that randomization represents the gold standard for the study of causation. The implicit picture of causality, then, is that a cause X is what make a difference to an an effect Y. To provide evidence that X causes (i.e. makes a difference to) Y, researchers should show that a surgically induced change in X (assuming possible counfounders are controlled for) makes a difference to Y.

A pair of vignettes by Lily Hu[11] raises a puzzle for the interventionist picture of causality (and for audit studies):

> AUDIT STUDY I: Skirt Interview - Two actors, one taken to be male and the other female, present identical resumes, answer interviewer questions identically, and affect the same tone, mannerisms, and general personality traits (as best as they can). The male actor also dons the same dress and wears the same facial makeup as the female actor; both actors wear skirts and facial makeup to their interviews. (p. 86)

> AUDIT STUDY II Confident Men and Mild-Mannered Women - Two actors, one taken to be male and the other female, present identical resumes and answer interviewer questions identically. To avoid confounding by perception of gender nonconforming status that may be triggered by, for example, setting identical styles of dress and facial makeup across the auditors, the social scientists look to make sure that both actors display traits that they take to be gender-conforming. For example, though the male actor presents as a confident and assertive candidate, the social scientists have the female actor portray as mild-mannered and demure, as a part of the effort to maintain what they take to be gender-conforming affect. (p. 87)

These are imaginary studies. The first audit study ensures that the two interviewees are perfectly identical except for whether they are

[11] Hu (2022), *Causation in the Social World*. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences, Chapter 2.

taken to be male or female. This is what we would expect the perfect randomized control trial to do: have two otherwise identical units that differ only by the social category of interest, in this case sex. But oddly, something in this study is off. By controlling for all the *intrinsic* causes, the experimenter failed to control for *extrinsic* (relational) causes. Whether someone is perceived as gender conforming or not depends on the interaction between attire and gender category. The second audit study attempts to contain the spillover effects of gender non-conforming behaviors. Thus, male and female interviewees are both portrayed as gender-conforming. But oddly, this study is also off. So many characteristics have been adjusted for to ensure the interviewees are both gender conforming. It is unclear whether the study can say anything about the causal role of sex.

The upshot here might be that what it means to intervene on sex is unclear. Until we know what sex is, it is unclear how we should intervene on sex. But can we know what sex is *before* we embark in a study of its causal role? Hu thinks we cannot.[12]. The same sort of considerations will likely apply to race.[13]

So, then, should we do away with audit studies? Hu thinks they provide powerful evidence of discrimination. They provide direct evidence of discrimination and indirectly evidence of the causal role of race/sex in society. For Hu, the claim of discrimination does not rest on a causal claim: it is the other way around.[14] Whether two resumes *should* receive the same callback rates is a normative judgment that is presupposed in the set up of the audit study itself. Once we agree with this moral judgment—and we might very well not agree—any deviation from the set moral standard is evidence of wrongful racial discrimination.[15]

[12] "This brings us to what I take to be the crux of the matter: whether an account of what sex is and an account of the causal role that sex plays in the social world can be disentangled at all." (p. 104)

[13] How, exactly?

[14] Hu writes, intriguingly: "an audit study need not make any claim to causation in order to substantiate a claim of discrimination, because discrimination is not an essentially causal notion" (p. 117).

[15] Whether two resumes are "matched" for the purpose of the audit study—resumes that should receive the same callback rates—is a moral question, and only when that question is settled, an audit study can provide evidence of wrongful discrimination.

Theories of Race

To ask whether race can play a causal role and how this causal role can be studied led us to ask a more fundamental question: What is race? We looked at two social constructionist accounts, one by Sally Haslanger[16] and the other by Chike Jeffers.[17]

For Haslanger, group *G* is racialized when:

(a) certain bodily features are taken to be evidence of a common ancestral geographical origin and are used to demarcate group *G*;

(b) these features take on social meanings and individuals in *G* are assigned a social position of subordination or privilege, where this social positioning is viewed as justified;

(c) conditions (a) and (b) play a role in placing individuals in the social hierarchy as subordinate or privileged.

[16] Haslanger (2019), Tracing the Sociopolitical Reality of Race, in *What is Race? Four Philosophical Views*, Oxford University Press.

[17] Jeffers (2019), Cultural Contructionism, in *What is Race? Four Philosophical Views*, Oxford University Press.

An interesting—albeit controversial—consequence of this account is that race is intimately tied with social hierarchy. For Haslanger, race should be distinguished from ethnicity. Ethnic groups are primarily demarcated by culture (language, customs, etc.), and the process by which they are situated in a social hierarchy is the *racialization* of the ethnic group. So, ethnic groups can exist without social hierarchy, but racial groups cannot. If social hierarchy is eliminated, then racial groups would also be eliminated.

Jeffers is, like Haslanger, a constructionist about race. But he draws an important distinction between *political* and *cultural* constructionism. He holds, like Haslanger, that we should move toward ending racialization insofar as it involves hierarchy and subordination. He also thinks that 'we ought to actively continue constructing races as cultural groups' (p. 58). For him, there is a distinctive cultural dimension to race which political constructionism overlook.[18]

We also examined two other views about race, biological race realism and race antirealism.

It is not uncommon to think of race as a risk factor in medical research. People of a certain race can be more at risk of developing a disease than people in other races. But if races are socially constructed or unreal, it is unclear how we can make sense of race-based differences in risk. If, instead, races are biologically real, this task would be easier. Spencer offers an argument that races are—in a qualified sense—biologically real.[19] The starting point of the argument is the racial classification of the Office of Management and Budget (OMB). In OMB race talk, Hispanics are not a race, but an ethnicity. Racial categories are divided into five groups: Asian; Black; American Indian; Native Hawaiian; White. If we can take OMB race talk as corresponding to ordinary race talk, what do ordinary people and experts alike *refer to* when they use racial categories? Do they refer to anything biologically real?[20] Spencer's answer is twofold. First, by studying patterns of allele frequencies, geneticists have found that human populations can be divided into five clusters: Africans, Eurasians, East Asians, Oceanians, and Native Americans. These human continental populations roughly correspond to the five OMB racial groups. Second, these clusters are biologically real. Spencer provides a number of criteria for an entity *e* to count as biologically real, roughly, the entity should play an explanatory or predictive role in a well-established biological theory. As it turns out, population geneticists are able to predict with high accuracy people's self-reported race based on an analysis of their genetic make-up. This predictive ability is evidence that race may well be biologically real.

The views considered so far all assume that race is real, either as a social reality (Haslanger, Jeffers) or as a biological reality (Spencer).

[18] Jeffers writes (pp. 64-65): 'three forms of cultural significance—racial consciousness itself as cultural, racial consciousness as facilitating new cultural developments, and racial consciousness as shaped by prior cultural developments—are key aspects of a proper account of the social construction of race, on my view. They are not central to standard political constructionism.'

[19] Spencer (2019), How to Be a Biological Racial Realist, in *What is Race?*

[20] Spencer is relying here on a *referentialist*, not a *descriptivist* theory of meaning.

But there are good reasons to think that race is neither socially nor biologically real. And if it is neither of these two, then race does not exist. This is Glasgow's position.[21]

[21] Glasgow (2019), Is race an Illusion of a (Very) Basic reality?, in *What is Race?*

Intermezzo

With a better understanding of what race might be, we should now go back to the original question. Can race be a cause and how should the causal effects of race be studied? Some scattered thoughts:

On Haslanger's account, race is the result of a *racialization process*. It is created and maintained by it. This process—roughly—picks up on bodily features (skin color, hair texture, etc.) and assigns them social meanings; people are then categorized and placed in a social hierarchy. So, on this view, racial discrimination is one way in which racialization unfolds. If this is right, what can audit studies tell us? They provide us information about the racialization process itself. Suppose an audit study shows people with Black sounding names receive, on average, fewer job callbacks than people with White sounding names. This difference in callback rates shows that people's names are one of the features picked up by the process of racialization. If the study did not show a difference in callback rates by name, this would tell us that racialization does not operate via names. It may operate in other ways, to be discovered. So, on Haslanger's account, while the existence of racial discrimination cannot be denied so long as races exist, the way in which racial discrimination operates can still be empirically investigated.

Suppose we run a plethora of audit studies that *all* visible features that could be picked up by the racialization process (accent, skin color, name sounding, hair texture, etc.) and we find no difference, on average, in a wide range of outcomes (loans, hiring, etc.). For Haslanger, these audit studies would show that racial discrimination does not exist; they would also show that race itself does not exist.[22]

[22] Could Hu agree with this conclusion? It seems so. Note that the conclusion assumes that race and racial discrimination cannot be disentangled one from the other.

But what to say about the causal role of race? On Hanslanger's account, race is not variable in a causal model that can be turned on and off. People's name, hair texture, skin color, etc. can be turned on and off. It is possible to design audit studies in which these features are manipulated. But this does not amount to manipulate race itself, which isn't an isolated variable. Race is the result of the entire process of racialization. So we should be wary of the limited information that audit studies give us.

If race is a not variable, how should it be represented in a formal causal model? To answer this question, the distinction between causes as triggers and causes as constraints is helpful.[23] Variables represent causes as triggers (they can be turned on and off, like a

[23] Dretske (1988), *Explaining behavior: Reasons in a world of causes*, MIT Press.

light switch). Constraints instead represent causes as structures of variables (think about how an electric circuit is wired). So should race be represented in causal model as a structure/constraint? If so, how?

Structures and Constraints as Causes

To better understand how structures can be causes, we read a much discussed paper on structural explanation by Haslanger. In it, she argues that we understand structural causes in terms of a part/whole relationship.[24] But this seems insufficient. While some part/whole relationships are explanatory and causal, others are not. How are we to distinguish between the two? Instead of thinking in part/whole terms, Ross prefers to think of structural causes as *constraints*.[25] The core of Ross's proposal: structures as causes should be understood as *constraints* on the available choices individuals can make. The constraints that a structure imposes on individual choices can be more or less stringent. The constraining force can be so definitive that it necessarily determines what individuals will do. More often than not, however, constraining causes will have an effect on the *probabilities* of possible outcomes, choices, and decisions.

One question in Ross's account is left open, though. If structures can sometimes act as causes by playing the role of constraints, how do they fit into the manipulability account of causation? Usually, single variables are manipulated: they are turned on or off. But structures do not seem to be just single variables. What would a manipulation or an intervention on a structure/constraint look like? To answer this question, we read a paper by Malinsky.[26]

To fix ideas, here is a simple structural equation model:

$$Y = \theta_1 X_1 + \theta_2 X_2$$

The model says that variables X_1 and X_2 are (potential) direct causes of the outcome variable Y. The strength of the causal dependency is given by the coefficient θ_1 and θ_2. Suppose that $\theta_1 = .4$ and $\theta_2 = -1.3$. The vector $(\theta_1, \theta_2) = (.4, -1.3)$ gives information about the *causal structure*. So, intervening on the structure means—at least in this example—to change the values assigned to the parameters θ_1 and θ_2, say to $(0, .87)$. Once the intervention is carried out, the counterfactual (causal) question can be asked: by manipulating the structure (that is, manipulating the values of the parameters), how would the value of the variables change as a result? This is rather abstract, but can be applied to concrete examples.[27]

[24] Recall the examples Haslanger uses: the treat in the ball; the grading curve; standing up when the Queen enters; the invisible foot; see Haslanger (2015), What Is a (Social) Structural Explanation? *Philosophical Studies*

[25] Ross (2023), What Is Social Structural Explanation? A Causal Account, *Nous*.

[26] Malinsky (2018), Intervening on Structure, *Synthese* 195:2295-2312.

[27] An interesting (and perhaps revealing) limitation of this approach is that structures with feedback loops cannot be modeled as causes. So, looks like the Invisible Foot example by Haslanger cannot be modeled as a cause. How does this observation apply to the causal role of race and gender? Does modeling their causal role require postulating feedback loops?

Coda

Let's now return to our earlier question. Should race be represented in a causal model as a structure/constraint? If so, how? Do structures as causes help to understand the causal role of race (or gender)? Here is a tentative responses. Following Ross, race and gender can be modeled as structures that constrain the choices of individuals. Following Malinsky, race and gender can be modeled as the parameters in a given causal causal structure. Let's develop a bit more the latter idea. Consider a simple causal model:

$$I = \theta_1 T + \theta_2 C,$$

where I stands for income level, T for hair texture and C for skin color. Say the coefficients θ_1 and θ_2 posits a correlation between hair texture and skin color, and income level: certain hair textures and skin colors are correlated with lower income levels, etc. Crucially, race or gender do not appear as variables. We can think of this graph, along with its coefficients, as modelling (part of) the process of racialization. Hair texture and skin color are one of those features picked up by racialization. Take a more complicated causal graph, with hundreds of variables and coefficients. A particular matrix of coefficients would correspond to a particular process of racialization. The key here is that race (or any other social category) never appears as variables in the causal model. Race is modeled as the entire matrix of coefficients.[28]

Does this formal model of race (or racialization) as a matrix of coefficient address Hu's concern about audit studies? Unfortunately, it seems not. Formally stated, Hu's concern amounts to the following observation: the matrix of coefficients is not invariant under changes of the values of the variables in the causal model. To see why, recall Hu's Skirt Interview vignette. Whether an applicant is perceived as male or female changes the causal effect of the variable 'skirt' on the likelihood of being hired. The causal effect of wearing a skirt onto the outcome – modeled formally by the coefficients – is not the same depending on the value of the other variable, the applicant's perceived sex. So, a single matrix of coefficients cannot be the same for all combinations of values of the variables in the causal model. Then, a concept such as 'race' picks out a very complicated set of phenomena, and it is perhaps modeled by sets of matrices of coefficients, one matrix for each permutation of the values of the variables.

[28] Interestingly, this formal model vindicates the conjecture that race can operate as a constraint. If the coefficients are the way they are, certain outcomes are rendered more likely than others. This constraints on what individual choices can accomplish.