Race Causality Discrimination – Measuring Discrimination

Marcello Di Bello - ASU - Fall 2023 - Week #10

After studying causality during the first few weeks of the seminar, we spent the last couple of weeks reading about the philosophy of race, in particular, social constructionism[1], biological race realism and antirealism.[2] This week we turn to racial discrimination, a paradigmatic instance of how race can play a causal role. How should we think of racial discrimination in light of the discussion about race and causality that we have had in the past few weeks?

[1] We looked at the account of race in Haslanger (2019), Tracing the Sociopolitical Reality of Race, in *What is Race? Four Philosophical Views*, Oxford University Press, as well as the account by Jeffers (2019), Cultural Contructionism, in *What is Race?*

[2] See Spencer (2019), How to Be a Biological Racial Realist, in *What is Race?* and Glasgow (2019), Is race an Illusion of a (Very) Basic reality?, in *What is Race?*

Causal inference and discrimination

We begin with the received view among quantitative social scientists about how racial discrimination should be studied and measured. We will rely on a report by the National Academy of Sciences.[3]

Following the work of Holland and Woodward,[4] causal claims can be translated into claims of counterfactual dependence between one variable and another: variable X (say, fire) is a cause of another variable Y (say, temperature) whenever had X taken a different value (say, the fire is extinguished instead of ongoing), Y would have taken a different value (say, the temperature is lower instead of higher).

Racial discrimination can be understood in the same counterfactual fashion: had the individual been of a different race (say, White instead of Black), the individual would have been differently treated (say, hired instead of not). If this counterfactual is true, the decision was racially discriminatory.

The received view is that establishing a counterfactual claim of this sort would make a compelling case of racial discrimination. The challenge for social scientists, then, would be to provide empirical evidence and quantitative methods that can establish the truth or falsity of this counterfactual claim. The problem is that it is very hard to establish such a counterfactual.[5]

[3] See National Academies of Sciences, Engineering, and Medicine (2004), *Measuring Racial Discrimination*, The National Academies Press, especially Chapter 5 and the Executive Summary.

[4] Holland (1986), Statistics and Causal Inference, *Journal of the American Statistical Society*, 81(396); Woodward (2003), *Making Things Happen: A Theory of Causal Explanation*, 2003, Oxford University Press.

[5] Not only that: it is not completely clear what the counterfactual could mean.

The report hints at another understanding of the nature of causation, in terms of mechanism rather than counterfactuals, although it treats the two as closely connected:

> [the] notion of describing mechanisms relates to what this report refers to as understanding the process whereby discrimination may be occurring ... This report views the effort to measure the unobserved counterfactual usually associated with experiments as necessarily being linked to a detailed understanding of the process. (pp. 81-82)

Let's now turn from theory to practice. How is racial discrimination

actually studied in the quantitative social sciences? It is studied in different ways, specifically:

1. Using randomized controlled experiments, such as laboratory studies, and field and audit studies;[6] and
2. using observational methods, such as natural experiments, statistical analyses of observational data (e.g. regression analyses)[7], and survey studies.

It is harder to make causal inferences from observational data than data from randomized control trials. Causal claims are warranted given laboratory and audit studies, but they are usually not warranted given observational data unless several assumptions are made.[8] The National Academy of Sciences notes:

> For purposes of causal inference, there is a hierarchy of approaches to data collection. As one moves from meticulously designed and executed laboratory experiments through the variety of studies based on observational data, increasingly strong assumptions are needed to support the claim that X "causes" Y. (p. 82)

Despite this hierarchy, however, laboratory and fields studies suffer from their own limitations. They tend to be narrower in scope, often limited to controlled laboratory settings that hardly capture complex social processes of which racial discrimination is presumably part. Another problem is that they tacitly assume what we might call an *on/off* conception of race.[9] If race could be turned on and off, it could be studied in randomized controlled studies.[10]

So, two limitations for the study of racial discrimination are pressing. First, it is difficult to draw causal inferences about race because randomized control trials that deploy race as a variable to be manipulated are hard to carry out. Second, when race (or, to be careful, its proxy) can be manipulated in controlled experimental settings, the resulting data will likely have limited applicability and fail to capture the complex ways in which race operates in society.

In the end, the National Academy of Sciences recommends a multi-disciplinary and multi-methodological approach:

> How is causality established in the absence of a perfectly designed and implemented experiment? It is possible to provide a stronger argument for causal inference by combining methods—from laboratory studies of proposed mechanisms, to field experiments demonstrating external validity, to natural experiments demonstrating policy relevance and efficacy. (p. 85)[11]

Key here is the aggregation of different lines of evidence that result from applying different research methodologies.[12]

[6] Can you think of examples? This method suffers from the problem of *external validity*. Can you say why?

[7] Can you think of examples? This method suffers from the *omitted variable bias problem*. Can you say why? How does randomization avoid this problem? See discussion on page 84.

[8] Can you give an example of a causal claim warranted by observational data given certain (which?) assumptions?

[9] Think of the story *The Sneetches* by Dr. Seuss: "In the story, one race of Sneetches is afforded certain privileges for having stars on their bellies, and the other race, lacking these markings, is denied those same privileges. There are, however, Star-On and Star-Off machines that can alter the belly and therefore the race of both Plain-Belly and Star-Belly Sneetches. Thanks to these machines, an individual Sneetch's racial status and various outcomes could be observed more than once, both as a Plain-Belly and a Star-Belly Sneetch." (p. 77) Further: "In *The Sneetches*, belly-based discrimination is evident in the society; the causal relationship between race and discrimination can be ascertained because stars can be placed on or removed from any belly by a machine, and multiple outcomes can be observed for a single Sneetch." (p. 78)

[10] In fact, race cannot simply be turned on and off. Is this claim supported by the accounts (which ones?) of race we have studied in past two weeks?

[11] Why does the report cite the example of lung cancer and smoking? What can we learn from that example according to the National Academy of Sciences?

[12] Even though—in theory—the National Academy favors a counterfactual approach to studying racial discrimination, this is less clear in practice as it recommends a richer array of methods.

Counterfactuals in the law

The counterfactual approach to racial discrimination is standard in the social sciences. The question is whether it is adequate.[13] Litigation about racial discrimination will sometimes appeal to counterfactuals of the form 'had the person been of a different race, etc.' For exmaple, in the case Students for Fair Admissions (SFFA) v. Harvard College, 600 US – (2023), plaintiff's expert Arcidiacono, professor of Economics, writes:

> Consider the example of an Asian-American applicant who is male, is not disadvantaged, and has other characteristics that result in a 25% chance of admission. Simply changing the race of the applicant to white–and leaving all his other characteristics the same–would increase his chance of admission to 36% ... Changing his race to African-American (again, leaving all other characteristics the same) would increase his chance of admission to 95%. (pp. 27-28)

How are we supposed to understand this counterfactual without knowing the broader causal processes? To see what the issue is, let's leave race for a moment.

Consider a case of religious employment discrimination. A plaintiff claims they were discriminated against in hiring because of their Catholic religion. Suppose the hiring decision has a disparate impact on those of religion $A = 1$ (say Catholics).[14] A lower percentage of Catholic applicants were hired. So, the plaintiff who is Catholic does have a *prima facie* claim of disparate impact. However, after controlling for educational attainment, the disparity disappears. Should the claim of religious discrimination be dismissed? It depends. There are different ways to model this scenario.

First, religion can be a factor affecting one's life choices, including education. Perhaps a (discriminatory?) process forces Catholics to attend less rigorous schools, which results in Catholics being less prepared for high paying jobs. This can be represented by this causal graph: $A \to Z \to Y$. Here, a causal process operates from A (religion) to Y (hiring) via the *mediator* variable Z (educational attainment). Hiring decisions (Y) are based on educational qualifications (Z) and one's religious (A) affects one's educational attainment (Z). On this interpretation, one's religion affiliation A causally affects the hiring decision Y, indirectly via Z. So the following counterfactual is true: had the job applicant been of a different religion (say, Protestant instead of Catholic), they would have been treated differently (say, they would have been hired).[15] Given the counterfactual account of racial discrimination, the Catholic applicant was discriminated against on the basis of religion.[16] Controlling for Z was therefore unwarranted: there is a causal influence of religion onto hiring (mediated by ed-

[13] See Barocas, Hardt and Narayanan (2023) *Causality*, especially section "Counterfactuals in the law", Chapter 5 of *Fairness and Machine Learning*, MIT Press.

[14] A claim of **disparate impact** discrimination in employment is established, as follows. *STEP I*: plaintiff makes a *prima facie* case showing that members of a protected group are disparately impacted (say they are hired for high paying positions at lower rates). *STEP II*: defendant (employer) responds that the employment practice has a job-related purpose ("business necessity defense"). *STEP III*: plaintiff establishes that the employment could have achieved the job-related purpose with an alternative employment practice with no disparate impact. The other legal doctrine of discrimination is **disparate treatment**. What is the difference with disparate impact?

[15] Whys it this counterfactual true?

[16] Whether this claim would stand up in a court of law is another matter. Also, in the example, there is already a discriminatory mechanism to the disadvantage of Catholics in college admissions. While there might be *educational* discrimination against Catholics, there might not necessarily be *hiring* discrimination against Catholics. These distinctions can be made via causal diagrams, but simple counterfactuals cannot capture them.

ucation), and controlling for education incorrectly eliminates this causal influence. Thus, it is not always recommended to control for additional variables, especially if they are *mediators* in a causal path.

But this scenario can be viewed differently. Suppose educational attainment affects one's religion. The more one is educated, the less likely one will profess a certain religion. Here the causal graph is different: $A \leftarrow Z \rightarrow Y$. On this interpretation, variable Z is a *confounder* because it is a common cause of the other two variables A and Y. It is not, like before, a mediator in a causal path from one variable to the other. If controlling for religion eliminates the disparity in hiring rates between Catholics and other religion groups, this would indeed show that there was no causal influence of religion onto the hiring decision. So there was no religious discrimination.

The upshot is this: an understanding of the larger causal processes at play affords us a better understanding of the ways in which religion plays a causal role. The same scenario could be interpreted in different ways: one warranted the claim of religious discrimination and the other did not. And this was so given the same observational data. The same conclusion will apply to race. The nature of race, its place and function in society, how it operates and interacts with other social categories, etc. are all controversial matters. So, claims of racial discrimination would require us to settle many empirical and normative claims about the larger causal processes in which race is implicated. If this right, a counterfactual approach to racial discrimination seems too narrow and limited in scope.

The causal model approach has its own limitations, though. Here are a couple worth considering:

- *Modularity*: Each node in a causal must be specified and is assumed to contribute a causal effect onto other variables. Other nodes will also contribute causal effects on their own and the overall causal effect onto a variable will be the result of the interaction of the different nodes. This modularity of effects might not always hold, specially in the case of race and gender.[17]
- *Individuality*: Individuals are the units of investigation.[18] It is unclear whether the model can capture structural group-level racial discrimination.[19]

[17] "Suppose that group membership is constructed from a set of social facts about the group and practices of individuals within the group. We might have some understanding of how these facts and practices constitutively identify group membership. But we may not have an understanding of how each factor individually interacts with each other factor, or whether such a decomposition is even possible" (p. 31)

[18] Why does the causal model approach rely on individuals as the units of investigation?

[19] This will be a topic of discussion in the next class.