Race Causality Discrimination – Is Race a Cause?

Marcello Di Bello - ASU - Fall 2023 - Week #4

Two weeks ago we looked at the Rubin model of causation as described by statistician Paul Holland[1]. Last week we turned to the philosophical literature, in particular, Woodward's manipulability theory of causation.[2]

The hope is that by now we have acquired a reasonably good sense of what causality consists in, at least from the perspective of the manipulability account. It is time to ask one of the central questions of this course, can RACE[3] be a cause? Holland thinks the answer should be negative, while others disagree.[4]

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

[2] James Woodward (2003), *Making Things Happen: A Theory of Causal Explanation*, 2003, Oxford University Press

[3] We use Holland's convention of referring to the variable 'race' simply as RACE.

[4] See Holland (2003), Race and Cause, *Research Report*, January 2003 RR-03-03, ETS Educational Testing Services and Marcellesi (2013), Is Race a Cause?, *Philosophy of Science*, 80(5).

Holland's argument that RACE cannot be a cause

Let's first recall Holland's account of causality. For him, the causal effect of t (relative to c) on unit u (measured by a response variable Y) is expressed by the following 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.[5]

[5] How does Woodward's account of causality differ from Holland's?

Let's see what happens when we apply this account of causality to RACE. The causal effect of race $RACE = b$ (compared to another race $RACE = w$) 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 $RACE = b$ 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:

> because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black. (p. 9)

Now, if RACE cannot be a causal variable, this raises the more general question of what can count as a causal variable. 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:

> The only rule I have is that if the variable *could be* a treatment in an experiment (even one that might be impossible to actually pull off due to ethical or practical issues), then the variable is probably a cause and correctly called a *causal variable*. (p. 9)

Surprisingly, this criterion of causality makes many variables non-causal. Test scores, age and gender, as well as race, no longer count as causal variables.[6]

[6] "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)

An Objection

Some will object that plenty of studies manipulate RACE as an experimental variable. For example, there are studies in which identical pairs of resumes are sent out to employers and the only variable that changes is RACE. Holland's response is that such studies merely change a manifestation or aspect of RACE, not RACE itself:

> These examples show, instead, how complex the manipulation of race really is ... In the resume studies, it was only the RACE on the resume that was changed—altering provided information—not the life experiences that accompany a resume in real life. Although not entirely irrelevant, this is a far cry from changing the race of a "real" individual. (p. 10)

A clarification is in order. Recall that Holland's account of causality focuses on causal effects, and not so much on causal mechanism.[7] Arguably, Holland is willing to admit that RACE can be part of a causal mechanism that explains a phenomenon or pattern in the data. What he objects to is that the causal effect of RACE can be studied experimentally.[8]

[7] He writes: "It should be clear, but often is not, that the language of causation is more precise when we are concerned with assessing effects than when we are concerned with either identifying causes or with proposing causal mechanisms. In the latter two cases, anything can be a cause, because we are just talking rather than doing. When we design an experiment, or other causal study, however, the only things that can qualify as causes are treatments or interventions." (p. 8)

[8] If this interpretation is right, it seems as though his argument is only about the limitations of experimental studies in examining RACE as a cause, not so much the imposisbility of RACE being a cause as such. Do you agree?

Discrimination, Biases and Causation

In the final part of his paper, Holland addresses a couple of lingering questions:

- How can we meaningfully talk about racial discrimination if RACE is not a cause, given that discrimination seems to require RACE to play a causal role?
- If RACE is not a cause, can we still study the causal effects of RACE in experimental studies, and if so how?[9]

[9] See Holland's extended discussion of biased test on pp. 13-19.

To both these questions, Holland responds that we should study the "statistical *interaction* of RACE with an appropriate difference in society" (p. 12). The study of these interactions can be useful to make claims about discrimination or biases.[10]

[10] What is the 'fantasy' that Holland describes on pp. 11 and 12? What is the point of that 'fantasy'?

Marcellesi against Holland

It is helpful to state Holland's argument against RACE as a cause more precisely. Marcellesi's reconstruction goes as follows:

> Step 1: Race is a necessary property of units. (This is an *assumption*.)
>
> Step 2: If a unit u is of RACE=b, then it is impossible for u to have been of another race RACE=w. (This follows from the definition of necessity in Step 1.)
>
> Step 3: Counterfactuals of the form 'Had unity u been of race RACE=w instead of RACE=b, then…' cannot be non-vacuously true. (This follows from Step 2 because the antecedent of the counterfactual would always be false.)
>
> Step 4: The causal effect of RACE is not defined in Holland's theory of causality. (This follows from Step 3 and Hollan's definition of causal effect.[11])
>
> Step 5: If x is a cause, then the causal effect of x must be defined in Holland's theory of causality. (This is an *assumption*.)
>
> Conclusion: RACE is not a cause. (This conclusion follows from Step 4, Step 5 and *modus tollens*.)

[11] See earlier the formula: $Y_b(u) - Y_w(u)$.

This argument is valid, but rest on two questionable assumptions: Step 1 and Step 4. So, unsurprisingly, Marcellesi attacks both of them. Why think that RACE is a necessary attribute (Step 1)? Why think that a well-defined causal effect in Holland's theory is a necessary requirement for being a cause (Step 5)?

Against Step 1, Marcellesi provides the following thought experiment:

> Consider the following hypothetical randomized experiment: assume that the race r_i of unit i is a function $r_i = f(b_i, e_i)$ of biological (b_i) and environmental (including social and cultural) factors ($\$e_i$). Imagine that values of b_i and e_i, and thus also of r_i, are randomly assigned to embryos 30 days after conception. The biological factors are assigned via genetic engineering, and the environmental factors are assigned by swapping embryos between mothers. (p. 655)

Even though this experiment is not practically possible—and it is not even ethical—it is in principle possible. If it is in principle possible, then RACE need not a necessary attribute of individuals, and counterfactual of the form 'Had unity u been of race RACE=w instead of RACE=b, then…' would be perfectly intelligible.[12]

Now, against Step 5, Marcellesi points out that there are plenty of cases of causal processes that are not modeled by Holland's theory:

[12] As Marcellesi admits (footnote 8, p. 655), if one thinks that race is determined by genealogical factors such as the identity of one's biological parents, then the imaginary experiment is not manipulating RACE. Is this a problem for Marcellesi's argument?

> Consider, for instance, the case of primary school performance: according to Holland himself, scholastic achievement in primary school cannot be treated as a cause of the choice of secondary school. . . there are very good reasons to think that how well a student does inprimary school has a causal effect on what secondary school she chooses to attend, for example, by determining what schools she is admitted to. The right conclusion to draw . . . seems to be that some genuine causes cannot be handled by [Holland's theory of causation] (p. 653)

In other words, Marcellesi thinks that examples like scholastic achievement show that Holland's theory of causation is incomplete since it cannot handle obvious cases of causation. Hence, Step 5 fails. It is not a necessary requirement of causality that the causal effects must always be defined in Holland's theory of causality.

Perceptions of race, not race itself

Marcellesi holds that race can be part of a causal explanation. Consider, for example, a society in which two (otherwise homogeneous) racial groups, A and B, are paid differently by an employer.[13]

[13] The assumption here is that the two groups are otherwise indistinguishable except for their race.

> What is the mechanism generating the wage gap in this society? What explains the fact that some A worker, call her w_A, receives wages 30% lower than those of a B worker, call her w_b, occupying an equivalent job? One straightforward answer is that w_A receives wages 30% lower than those of w_b because she is an A and because the employer believes the work of A's to be worth 30% less than that of Bs. In other words, the fact that w_A is an A, together with the employer's belief about the relative worth of the work of As, is the cause of her receiving wages 30% lower than those of w_B (p. 656)

This is a rather simple and plausible explanation. Nothing surprising goes on here. However, one might object that RACE itself does not cause lower wages. Rather, what causes lower wages is what employers believe about people's races. So, the causal mechanism seems to comprise *perceptions* of race, not race itself.

The obvious response is that perceptions of race must be caused by race and thus race would—ultimately—be the cause. But what if perceptions of race are not caused by race?

> if perceptions of race are not caused by race but, rather, by features the instantiation of which is merely correlated with race, then it is not clear that discrimination on the basis of these perceptions is properly described as racial discrimination.

In other words, if the employer's decisions are not caused by race, but rather, by the prospective employee's skin color (a feature that correlates with, but is not identical with, race), it is unclear how the employer's behavior can ever count as racial discrimination.[14]

[14] Presumably, if a theory of causality rules out racial discrimination, it cannot be a good theory. Do you think that Holland's discussion of racial discrimination addresses this worry?