Is race a cause?

Alexandre Marcellesi

Draft of 11/03/2012
To be published in revised form in *Philosophy of Science*

Comments welcome (amarcellesi@ucsd.edu)

Abstract

Advocates of the counterfactual approach to causal inference argue that race isn't a cause. I object that their argument is invalid and that its key premise is unwarranted. I also criticize the criterion, which I call 'Holland's rule', the counterfactual approach relies on to distinguish causes from non-causes. Finally, I argue that racial discrimination cannot be causally explained unless one assumes race to be a cause. I conclude that the view that race is not a cause lacks support and that there are good reasons to adopt the opposite view that race is a cause.

1 Introduction

Scientists in many disciplines (economics, epidemiology, etc.) routinely treat race as a cause. Economists who study labor market discrimination, for instance, commonly build models involving race as an independent variable and give a causal interpretation of the coefficient attached to it.[1]

Are scientists who treat race as a cause fundamentally confused? Do policies based on their conclusions rest on shoddy evidence? This is what leading advocates of the counterfactual approach to causal inference (henceforth 'CFA') claim, arguing that since race is an "immutable characteristic" of individuals, one cannot coherently treat it as a cause.

After a brief introduction to the CFA (§2), I present the argument against race being a cause (§3). I then raise two objections to it (§4) and proceed to sketch a positive argument for race being a cause (§5). I conclude that advocates of the CFA lack justification for denying race the status of cause, and that there are good reasons to adopt the opposite view that race is a cause (§6).

2 The counterfactual approach

The CFA, first introduced by Rubin (1974), is the dominant approach to causal inference in statistics and in many social sciences. It has roots in the work of Fisher and Neyman on agricultural experiments.

[1] See e.g. (Kahn and Sherer, 1988) for a classic example that is representative of many studies in labor economics.

When only one cause is considered, counterfactual causal models essentially have the following components:[2]

- A population of units $i \in U$
- A binary causal exposure variable D taking value $d_i = 1$ when i is exposed to the cause (is in the 'treatment' state) and $d_i = 0$ when i is not (is in the 'control' state).
- Two potential outcome variables Y^1 and Y^0, where y_i^1 represents the value of the effect for i when i is exposed to the cause and y_i^0, the value of the effect for i when i is not exposed to the cause.

The individual-level causal effect (ICE) of D for i is typically defined as follows:

$$\delta_i = y_i^1 - y_i^0$$

This causal effect is equal to the difference between the value of the effect when i is exposed to the cause and the value of the effect when i is not. Since a given unit cannot be both exposed to the cause and not exposed to it at once, only one of y_i^1 and y_i^0 can be observed for any unit. If i is exposed to the cause, the value of y_i^1 is observable while the value of y_i^0 is counterfactual: It is the value the effect *would* have taken had i not been exposed to the cause; hence the name of the approach. Because only one of y_i^1 and y_i^0 can be observed, δ_i cannot be observed either. Holland dubs this the "fundamental problem of causal inference" (1986, 947).

There are various solutions to this problem, both in experimental and in observational contexts. These solutions provide techniques for estimating the ICE and other causal effects, or parameters, built upon it. My concern here is not with the problems that race might raise for the application of these estimation techniques.[3] It is, rather, with the problems that race allegedly raises for the very definition of causal effects, and of the ICE in particular.

3 The argument against race being a cause

The argument developed by leading advocates of the CFA against race being a cause can be reconstructed as follows:

1. Race is a necessary property of units
2. If a unit is of race r, then it is impossible for it to have been of another race r' (from 1)
3. Counterfactuals of the form 'Had i been of race r' instead of r, then...' cannot be (non-vacuously) true (from 2).

[2] I adopt the terminology and notation from (Morgan and Winship, 2007).

[3] Rubin (1986; 2011) argues that estimating the causal effects of race is difficult enough to warrant its dismissal as a cause. I agree with Heckman (2005) that arguments of this kind conflate definition and estimation: That it is difficult to estimate the causal effect of race does not warrant the conclusion that it is not a cause.

4. The ICE of race is undefined (from 3 and the definition of ICE).

∴ Race is not a cause (from 4).

Let me illustrate this argument. Assume that there are only two races, that D represents race, and that $d_i = 1$ when i is White and $d_i = 0$ when i is Black. To say that race is a necessary property, "immutable characteristic" (Greiner and Rubin, 2011), or "attribute" (Holland, 1986, 955) of units is to say that if $d_i = 1$ (resp. 0), then it could not have been the case that $d_i = 0$ (resp. 1). Because this is so, counterfactuals of the form 'Had it been the case that $d_i = 0$ instead of $d_i = 1$, then the value of Y^0 for i would have been y_i^0' cannot be non-vacuously true when $d_i = 1$. Because no such counterfactual can be non-vacuously true, the causal effect of race is undefined, and this regardless of what effect the variables Y^1 and Y^0 represent (wages, education, etc.).[4]

In Holland's words, "attributes of units [like race] are not the types of variables that lend themselves to *plausible states* of counterfactuality." (2003, 14, emphasis original) He adds: "Because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black." (ibid.) And because the causal effect of race cannot be defined unless there is a non-vacuously true answer to such a counterfactual query, Holland concludes that race is not a cause.

The consequences of this view are important. If race is not a cause, then as Greiner and Rubin point out, "attempts to infer the causal effects of such traits [as race] are incoherent." (2011, 775) Holland goes further by claiming that, "Attributing cause to RACE is merely confusing and unhelpful in an area where scientific study is already difficult" and that, "Obscuring [the topics of discrimination and bias] with simplistic calculations that do not attend to the proper role of RACE in a causal study helps no one." (2003, 24)

So, do the many scientists who treat race as a cause waste time and resources on incoherent studies that only obscure important topics like racial discrimination? I do not believe so and develop two objections to the argument against race being a cause.

4 Against the argument against race being a cause

4.1 The argument is invalid

The most straightforward objection to the argument presented in §2 is that its conclusion does not follow from its premises. The fact that the ICE of race is undefined only entails that race is not a cause if the following premise is added to the argument:

$4'$. For all x, if x is a cause, then its ICE is defined.

[4] The same point applies mutatis mutandis to other causal effects defined in the CFA, e.g. the average causal effect defined over U as $E[Y^1] - E[Y^0]$. Because ICE is the fundamental causal effect for the CFA, however, I focus on it in the present paper.

If one adds this premise, then the argument is valid. There are good reasons, however, to believe that this premise is false, i.e. there are good reasons to think that some genuine causes cannot be handled by the CFA.

Holland himself claims, for instance, that scholastic achievement in primary school cannot be treated as a cause of the choice of secondary school by the CFA (1986, 955). Setting aside the question of the justification for this claim, the right conclusion to draw here is not that scholastic achievement is not a cause of school choice: There are good reasons to think that how well a student does in primary school has an effect on what secondary school she chooses to attend (e.g. by determining what schools she's admitted to). Rather, the conclusion to draw is that some genuine causes cannot be handled by the CFA, and so that premise $4'$ is false.

This conclusion is bolstered by the existence of frameworks for causal inference, e.g. Ragin's qualitative comparative analysis framework (1987), that do not rely on counterfactuals to define causal effects and which can thus treat variables whose ICE is undefined as causes.

4.2 Why believe premise 1?

How do advocates of the CFA justify the claim that race is an attribute, i.e. a necessary property, of units?[5] Their justification for this claim derives entirely from an application of what I will call 'Holland's rule' (or 'HR'). According to HR,

> 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 [. . .] correctly called a *causal variable*. (2003, 9, emphasis original)

It is important to note that, for Holland, attributes and causal variables form a partition of the set of properties of a unit: If a property is not a causal variable, then it is an attribute. Holland claims that race could not be a treatment in an experiment and, applying HR, he thus concludes that it is not a causal variable but, rather, an attribute (ibid.).[6] Greiner and Rubin agree and invoke "the impossibility of manipulating such traits [as race] in a way analogous to administering a treatment in a randomized experiment" (2011, 775) as one of the sources of the incoherence of studies purporting to estimate the effect of race.

There are two important problems with HR. First, it is the wrong rule for advocates of the CFA to follow. According to the CFA, for the ICE of D on i to be defined, there must be some counterfactual state in which i is not exposed to D, assuming that i actually was exposed to D.

[5] Glymour has objected to Holland that, "If counterparts [in the sense of (Lewis, 1968)] are conceivable – and why not? – then counterfactuals that violate identity conditions are intelligible, and if counterfactuals are intelligible, then causal relations are as well." (Glymour, 1986) Holland, however, can answer this objection by saying that the problem with attributes is not that they engender counterfactuals which violate identity conditions, but that they engender counterfactuals with impossible antecedents. In other words, Holland could answer that though counterparts are conceivable, no counterpart of a White unit can be, e.g., Black. Because race is a necessary property, all counterparts of a White unit also are White.

[6] Note that Holland's argument is fallacious given the way HR is formulated: It denies the antecedent of HR and infers the negation of its consequent. I'm here adopting a charitable reading according to which it is *necessary* for a property of units to be a causal variable that it be a treatment in some possible experiment.

In other words, it must be possible for i not to have been exposed to D. But why think that the possibility of such a state requires the possibility of an experiment resulting in it being the case that i is not exposed to D? To hold this view is to hold the implausible view that it is possible that p only if it is possible for there to be an experiment of the right kind resulting in it being the case that p. The right slogan for the CFA thus isn't "No causation without [some hypothetical experimental] manipulation" (Holland, 1986, 959) but, rather, 'No causation without counterfactual states'. This slogan is less catchy but more faithful to the way the CFA defines causal effects (e.g. the ICE).

One might object that HR was intended by Holland not as a strict rule but as a heuristic. It is true that he prefaces his presentation of HR by saying that, "There is no cut-and-dried rule for deciding which variables in a study are causal and which are not." (2003, 9) It should be noted, however, that despite this caveat, Holland *does* apply HR as a "cut-and-dried" rule, since he takes the supposed violation of HR by race to be sufficient to establish the conclusion that race is an attribute and so cannot be a cause (op. cit., 10). It should also be noted that HR fares no better as a heuristic rule than it does as a strict rule. I have claimed above that the possibility of an experiment resulting in i not being exposed to D is not necessary for it to be possible that i is not exposed to D. If so, however, then there is no reason to take the inconceivability of such an experiment to be a reliable guide to the impossibility of a state in which i is not exposed to D.

The second issue is that HR is vague – What kinds of experiments are admissible? What does it mean to say that a variable *could be* a treatment in an experiment? – and that, as a result, it is unclear that it is genuinely impossible for there to be an experiment in which race is the treatment. Indeed, let me argue against this impossibility claim by describing a hypothetical randomized experiment in which race is the treatment:[7] Assume that the race r_i of i is a function $r_i = f(b_i, e_i)$ of biological (b_i) and environmental (including social and cultural) factors (e_i).[8] 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.[9]

This experiment has not been carried out, is morally objectionable, and *might* be practically impossible given present science and technology. But this does not mean that this experiment is impossible. Indeed, the experiment described seems to be nomologically possible, i.e. carrying it out would not seem to require the violation of any laws of nature. This experiment also clarifies what the antecedents of counterfactuals of the form 'Had i been of race r' instead of r, then...' claim. The race of i would have been different just in case i had been randomly assigned a combination of values of b_i and e_i giving rise to a value r' of r_i that differs from its actual value r.

[7] Note that HR does not require the relevant hypothetical experiments to be randomized. I am offering more than is required here.

[8] What the relative weights of b_i and e_i are is no concern of mine. If you think that race is entirely determined by biological factors, then give zero weight to e_i; and if you think that race is entirely determined by environmental (including social) factors, then give zero weight to b_i.

[9] Note that this experiment will not work if, among the biological factors represented by b_i, are 'genealogical' properties of i (e.g. who i's parents are). Thus, if you think that races are biological groups unified by genealogical relations (see e.g. Hardimon 2012), then you should think that the experiment described above does not randomly assign race.

It thus seems that, despite what Holland and Greiner and Rubin assume, it is possible for race to be a treatment in an experiment, even a randomized experiment, and so it is not the case that race violates HR. Even if HR was the right rule for advocates of the CFA to follow (a view I have argued against), then, its application to the case of race would not support the claim that race is an attribute of units rather than a causal variable, i.e. it would not support premise 1. So, not only is the argument against race being a cause invalid, as I have argued in §4.1, but its key premise also lacks proper support.

5 A positive argument for race being a cause: Explaining racial discrimination

Consider an imaginary society in which there are two exclusive and exhaustive racial groups, A and B. Assume that in this society there is a wage gap between As and Bs: As receive wages that are uniformly 30% lower than the wages received by Bs occupying equivalent jobs. Assume, further, that all the units in the population, be they A or B, are perfectly homogeneous regarding the causes of wages (other than, possibly, race), e.g. they received the same degree from the same school, they have the same number of years of experience, they have the same IT skills, they have the same interpersonal skills, they work equally hard, they have the same preferences regarding wages, etc. Assume, finally, that there is only one employer in this society, and that this employer fixes the wages of the workers hired.

What is the mechanism generating this wage gap, i.e. what causally explains the fact that As receive wages that are 30% lower than those of Bs? The seemingly obvious answer is that As receive lower wages precisely because they are As and because the employer believes that the work of As is worth 30% less than that of Bs. In other words, what causes As to receive lower wages is the fact that they are As combined with the fact that the employer believes the work of As to be worth 30% less than that of Bs.

This commonsensical causal explanation is unavailable to somebody who claims that race is not a cause. If being an A is not a cause, then being an A cannot, in combination with the employer's belief about the worth of the work of As, cause one to receive lower wages. But, then, what causally explains the wage gap between As and Bs? Let me consider two alternative ways one might answer this question below.[10]

The first alternative answer, defended by Holland (2003), consists in claiming that what causally explains the wage gap is not the racial difference between As and Bs but, rather, the (racially) discriminatory nature of the society I described. This answer, however, faces an immediate difficulty. For advocates of the CFA, 'being discriminatory' must satisfy HR in order for the discriminatory nature of society to be a cause of the wage gap. In other words, it must be possible for 'being discriminatory' to be assigned as a treatment to societies in some experiment. Is such an experiment

[10] I leave aside two implausible solutions: First, the solution which consists in claiming that the wage gap is a brute fact, i.e. has no causal explanation. Second, the solution which consists in claiming that a society of the kind I've described is impossible, and that As and Bs must differ in some respect other than their race in order for the wage gap to arise.

possible?

Holland attempts to justify his claim that it is by describing "a *parallel world* [...] in which things are so different that what we recognize in our own world as racial discrimination does not exist in this other world." (2003, 16, emphasis original) Though Holland attempts to further flesh out this "little fantasy" (ibid.), his description falls far short of a precise description of a hypothetical experiment. He does not specify, for instance, the hypothetical experimental manipulations involved in making a society discriminatory.[11]

The claim that 'being discriminatory' satisfies HR, and so might be a cause of the wage gap between As and Bs, thus lacks proper justification while, as I argued in §4.2, there are good reasons to think that race does satisfy HR. If HR is the right rule for advocates of the CFA to follow, then, there does not seem to be any good reason to favor Holland's alternative explanation of the wage gap over the commonsensical explanation I presented above. And if, as I argued in §4.1, HR is not the right rule for advocates of the CFA to follow, then one can simply object to Holland that, absent an account of what it means exactly for a society to be discriminatory, his proposed explanation is little more than a vague suggestion while the commonsensical explanation given above clearly identifies a mechanism that is sufficient to generate the wage gap between As and Bs. Whether HR is the right rule for advocates of the CFA to follow, then, there are good reasons to favor the commonsensical explanation over Holland's alternative.

The second alternative answer, defended by Greiner and Rubin (2011), among others, claims that what causes As to receive lower wages is not their race in combination with the employer's belief regarding the worth of their work, but the perception of their race by the employer in combination with this same belief. There are several problems with this answer. I here examine three.

First, in the imaginary case at hand, it is simple enough to pin down who's perception it is that's relevant to explaining the wage gap, since there is only one employer. But what if there were many employers, and what if the wages of As were on average, rather than uniformly, 30% lower than those of Bs? Who's perception would then be relevant? The collective perception of all the employers? Or the collective perception of only those employers who falsely believe the work of As to be worth less than that of Bs? If one is to appeal to perceptions of race to explain any real wage gap between racial groups, then one needs answers to these questions. Greiner and Rubin themselves point out the difficulty of answering these questions as one limitation of this approach (ibid., 783-784). And the problem is more severe even when one considers studies of the effect of race on education or access to health care: What is the proper interpretation in terms of perceptions of race of the causal effects estimated by these studies?

Second, if the move to perceptions is legitimate in the case of race, then why not adopt it for other properties of units? Why not think that *perceptions* of education or work experience, rather than education or work experience, are what's causally relevant to an individual's wages? The move from race to perceptions of race seems ad hoc and motivated entirely by the assumption, which I

[11] It seems that, in order to describe such hypothetical experimental manipulations, one would first have to pin down what it means for a society to be (racially) discriminatory, something Holland does not do.

have argued to be mistaken in §4, that race cannot be a cause according to the CFA.

Third, and this is the most pressing problem, what causes the employer in the imaginary society I have described to perceive A workers to be As? If race is not a cause, then one cannot claim that the cause at work is the fact that As are of race A. Leaving aside the implausible claim that perceptions of race are uncaused, the most plausible solution seems to be to claim that what causes the employer to perceive A workers to be As is the perception of some set of features F the presence of which is strongly correlated with being a A. This solution faces a dilemma. Either the features in set F constitute what it is to be an A, in which case being an A is, after all, a cause of the employer's perception of As as As. Or the features in F do not constitute what it means to be a A.

In this latter case, the belief the employer must have in order for the wage gap to appear is not the belief that the work of As is worth less than that of Bs, but the belief that the work of units exemplifying features F is worth less than the work of units which do not exemplify these features. If this is so, then describing the discrimination against As as *racial* discrimination is inappropriate: As are discriminated against not on the basis of their race but on the basis of features that happen to be strongly correlated with being an A. More generally, this solution amounts to denying that there can be genuinely *racial* (direct) discrimination, i.e. "*differential treatment on the basis of race* that disadvantages a racial group", as a panel of the US National Research Council defines it (Blank et al., 2004, 39, emphasis original).

Neither the alternative explanation defended by Holland nor that defended by Greiner and Rubin thus seem as satisfactory as the commonsensical explanation presented at the beginning of this section, and which assumes race to be among the causes of the wage gap between As and Bs. This provides some support for the claim that one needs to assume race to be a cause in order to explain racial discrimination. Of course, the explanations offered by Holland and Greiner and Rubin, though they are the most prominent alternatives in the literature (especially the latter), do not exhaust the space of possible alternatives to the commonsensical explanation of the wage gap. This is why, in the introduction to this paper, I described the discussion in the present section as *sketching* an argument for race being a cause.

6 Conclusion

Are the attempts of labor economists to infer the causal effect of race on, e.g., wages "incoherent", as Greiner and Rubin (2011, 775) claim? Is it the case that "Attributing cause to RACE is merely confusing and unhelpful", as Holland (2003, 24) maintains? I have here argued that there is no reason to think these claims to be true.

First, the argument advanced by advocates of the CFA against race being a cause is invalid. Second, its key premise, that race is an attribute of units, is not justified by the application of Holland's rule, a rule that advocates of the CFA should reject anyway. Third, there are good reasons to think that explaining racial discrimination requires one to treat race as a cause.

I have said nothing up to now about debates in the philosophy of race. The view defended in this paper bears on these debates in the following way: Whatever concept of race one thinks is fit for use by labor economists studying racial discrimination, one's account of this concept should imply that races can be causes.

The debate over the causal status of race examined in this paper also gives a useful example of a case where philosophers of science can, and should, contribute to clarifying the debate and critically examine the assumption made by the scientists involved. This is what I have tried to do above.

Acknowledgments

I wish to thank Craig Callender, Nancy Cartwright, Michael Hardimon, Gil Hertsthen, and Chris Wüthrich for comments on earlier drafts. I also wish to thank members of the audience at the UCSD Graduate Philosophy Colloquium, in particular Amy Berg, Dan Burnston, Joyce Havstad, Nat Jacobs, Chris Pariso, and Adam Streed.

References

Blank, Rebecca, Dabady, Marilyn, and Citro, Constance (eds.). 2004. *Measuring Racial Discrimination.* Panel on Methods for Assessing Discrimination. Washington, D.C.: The National Academies Press.

Glymour, Clark. 1986. "Comment: Statistics and Metaphysics." *Journal of the American Statistical Association* 81:964–966.

Greiner, James and Rubin, Donald. 2011. "Causal effects of perceived immutable characteristics." *The Review of Economics and Statistics* 93:775–785.

Hardimon, Michael. 2012. "The Idea of Scientific Concept of Race." *Journal of Philosophical Research* 37:249–282.

Heckman, James. 2005. "Rejoinder: Response to Sobel." *Sociological Methodology* 35:135–150.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

—. 2003. "Causation and Race." Technical Report RR-03-03, Educational Testing Services.

Kahn, Lawrence and Sherer, Peter. 1988. "Racial Differences in Professional Basketball Players' Compensation." *Journal of Labor Economics* 6:40–61.

Lewis, David. 1968. "Counterpart Theory and Quantified Modal Logic." *Journal of Philosophy* 65:113–26.

Morgan, Stephen and Winship, Christopher. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Ragin, Charles. 1987. *The Comparative Method*. University of California Press.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

—. 1986. "Comment: Which ifs have causal answers?" *Journal of the American Statistical Association* 81:961–962.