Is Race a Cause?

Alexandre Marcellesi*†

Advocates of the counterfactual approach to causal inference argue that race is not a cause, and this despite the fact that it is commonly treated as such by scientists in many disciplines. I object that their argument is unsound since two of its premises are false. I also sketch an argument to the effect that racial discrimination cannot be explained unless one assumes race to be 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, often build models involving race as an independent variable and interpret estimates of the coefficient attached to it as estimates of the causal effect of race. This practice conflicts with the view held by leading advocates of the counterfactual approach to causal inference (CFA) who argue that, since race is a necessary property of individuals, one cannot coherently treat it as a cause.

Important issues hang on the outcome of this debate between practitioners and theorists of causal inference. If race is not a cause, then the coefficients attached to variables representing race cannot represent the causal effect of race. But then what, if anything, do they represent? And if these coefficients cannot represent the causal effect of race, then is it legitimate to use data on race to estimate them? Should studies that purport to measure the causal effect of race (e.g., on earnings or on access to health care) be funded? And should social and health policies be based on results from such studies?

*To contact the author, please write to: Department of Philosophy, University of California, San Diego, 9500 Gilman Drive 0119, La Jolla, CA 92093-0119; e-mail: amarcellesi@ucsd.edu.

†I thank Craig Callender, Nancy Cartwright, Michael Hardimon, Gil Hertshten, and Chris Wüthrich for comments on earlier drafts. I also thank audiences at the PSA 2012 and at the University of California, San Diego, Graduate Philosophy Colloquium. Research for this article was supported by the God's Order, Man's Order, and the Order of Nature project (Templeton Foundation).

Philosophy of Science, 80 (December 2013) pp. 650–659. 0031-8248/2013/8005-0011$10.00
Copyright 2013 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/673721 Published online by Cambridge University Press

After a brief introduction to the CFA (sec. 2), I present the argument against race being a cause (sec. 3). I then raise objections against two of its premises (sec. 4) and sketch a positive argument for race being a cause (sec. 5).

2. The Counterfactual Approach. The CFA, developed primarily by Rubin (see, e.g., Rubin 1974), is the dominant approach to causal inference in statistics and in many social and biomedical sciences. It has roots in the work of Fisher and Neyman on agricultural experiments.

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

- A population of units $i \in U$.
- A binary causal exposure variable D taking the 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:

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

The ICE 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. 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 one can build from it. My concern here is not with the problems that race might raise for the application of these estimation techniques. 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 advocates of the CFA against race being a cause can be reconstructed as follows:

https://doi.org/10.1086/673721 Published online by Cambridge University Press

1. Race is a necessary property of units.
2. If i is of race r, then it is impossible for i 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 (nonvacuously) true (from 2).
4. The ICE of race is undefined (from 3 and the definition of ICE).
5. For all x, if x is a cause, then its ICE is defined.

∴ Race is not a cause (from 4 and 5).[1]

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. Leading advocates of the CFA, such as Rubin and Holland, hold that race is a necessary property, "immutable characteristic" (Greiner and Rubin 2011), or "attribute" (Holland 1986, 955) of units. To say that race is a necessary property of units is to say that if $d_i = 1$ (respectively, 0), then it could not have been the case that $d_i = 0$ (respectively, 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 nonvacuously true when $d_i = 1$ (and conversely when $d_i = 0$). In Holland's words, "attributes of units [e.g., race] are not the types of variables that lend themselves to *plausible states* of counterfactuality" (2003, 14).[2] Because no such counterfactual can be nonvacuously true, however, the ICE of race is undefined, and this, regardless of what effect the potential outcome variables Y^1 and Y^0 represent (earnings, education, etc.).[3] And since the ICE of race is undefined, 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" and that "obscuring [the topics of discrimination and bias] with simplistic cal-

1. Note that the argument, thus reconstructed, is immune to the objection raised by Glymour (1986) against Holland (1986). Glymour objects that, "If counterparts [in Lewis's sense] are conceivable—and why not?—then counterfactuals that violate identity conditions are intelligible, and if counterfactuals are intelligible, then causal relations are as well" (1986, 966). If the problem with race is that it is a necessary property of individuals, however, then whether one favors transworld identity—as Holland implicitly does—or counterpart theory is irrelevant, and no appeal to the latter will help. If being of race r is a necessary property of i, then all the counterparts of i also are of race r, and so counterfactuals of the form 'Had i been of race r' instead of r, then . . .' have impossible antecedents and cannot be (nonvacuously) true.

2. Holland 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" (2003, 14).

3. The same point applies mutatis mutandis to other causal effects defined in the CFA, e.g., the average causal effect defined over U as $E[Y^1] - E[Y^0]$.

https://doi.org/10.1086/673721 Published online by Cambridge University Press

culations 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 now turn to two objections to the argument against race being a cause.

4. Against the Argument against Race Being a Cause.

4.1. Why Believe Premise 5? According to premise 5 in the argument against race being a cause, having a well-defined ICE is a necessary condition for being a cause. To believe this premise is to believe that every cause can be handled by the CFA. There are good reasons, however, to doubt this claim.

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 by the CFA because its ICE is undefined (1986, 955).[4] Assuming for a minute that Holland is correct in his assessment, the right conclusion to draw here does not seem to be that scholastic achievement is not a cause of school choice. This is so because there are very good reasons to think that how well a student does in primary 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, rather, seems to be that some genuine causes cannot be handled by the CFA and, therefore, that having a well-defined ICE is not necessary to be a cause.[5] This conclusion is bolstered by the existence of frameworks for causal inference, for example, Ragin's qualitative comparative analysis framework (1987), that do not rely on counterfactuals to define causal effects and that can thus treat properties whose ICE is undefined as causes.

4. Holland holds this view because he thinks that "it is difficult to conceive of how scholastic achievement could be a treatment in an experiment" (1986, 955) and because, as a result, he thinks that scholastic achievement, like race, does not lend itself to "plausible states of counterfactuality." Although Holland's reasoning is faulty—because it relies on a principle that advocates of the CFA should reject, as I will argue in sec. 4.2—let me assume here, for the sake of argument, that the conclusion he reaches is true.

5. The same conclusion seems warranted in the case of the ICE of the age at which a student starts school on her first grade test scores, a causal effect econometricians Angrist and Pischke dismiss as "impossible to interpret" (2009, 7) in the CFA and as giving rise to "a fundamentally unidentified question" (5). Because there are good reasons to think that the age at which a student starts school has a causal effect on her first grade test scores, the fact that this causal effect is "impossible to interpret" in the CFA suggests that there are genuine causes the CFA cannot handle.

https://doi.org/10.1086/673721 Published online by Cambridge University Press

4.2. Why Believe Premise 1? Why should one believe the claim that race is a necessary property, or attribute (in Holland's terms), of units? How do advocates of the CFA justify this claim? Their justification derives entirely from an application of what I will call 'Holland's rule' (HR). As Holland originally formulates it, HR states 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 . . . correctly called a *causal variable*" (2003, 9). It is important to note that, for Holland, attributes and causal variables form a partition of the set of properties of a unit: a property is an attribute if and only if it is not a causal variable.[6] Holland claims that race could not be a treatment in an experiment and, applying HR, concludes that it is not a causal variable but, rather, an attribute or necessary property (9). As should be obvious, Holland's argument is fallacious given the way he formulates HR: it denies the antecedent of HR and infers the negation of its consequent. I will here adopt a charitable reading according to which being a treatment in some possible experiment is both sufficient and necessary for a property to be a causal variable. The proper formulation of HR—and the one I will discuss below—is thus as follows:

> (**HR**) A property is a causal variable if and only if it could be a treatment in an experiment (even one that might be impossible to actually pull off due to ethical or practical issues).

Greiner and Rubin agree with Holland's line of argument 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 the main source of the incoherence of studies purporting to estimate the causal effect of race.

Given the biconditional formulation of HR above, Holland's argument to the effect that race is an attribute is valid. There are, however, two issues with HR. First, it is the wrong rule for advocates of the CFA to follow; that is, advocates of the CFA should see HR as false. 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 is exposed to D. 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—where p is of the form 'i is exposed (respectively, not exposed) to D'—only if it is

6. Note that I use 'property' and 'variable' as synonyms in this article (as does Holland in his writings). This is without consequences for the arguments developed.

https://doi.org/10.1086/673721 Published online by Cambridge University Press

possible for there to be an experiment resulting in it being the case that *p*. The right slogan for the CFA thus is not "No causation without [some possible 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 Holland prefaces his discussion 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). But note that, despite this caveat, he 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 is not a causal variable (10). It should also be noted that HR fares no better as a heuristic 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 with HR is that it is vague 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. 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).[7] 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.

This experiment has not been carried out, is morally objectionable, and is (presumably) practically impossible given present science and technology. But, according to Holland himself, this does not mean that this experiment is impossible. HR, however, does not give one any more guidance regarding what it means for an experiment to be possible. I take it to be obvious that this experiment is logically possible. This experiment also seems to be nomologically possible; that is, it does not seem that carrying it out would require the violation of any laws of nature. Is this experiment also conceptually possible? Not if your favorite concept of race implies that values of b_i and e_i, that is, biological and environmental factors, are not sufficient to determine an individual's race.[8] But if your favorite concept of race has this

7. You can set the relative weights of b_i and e_i however you like. This setup is intended to be as neutral as possible between concepts of race.

8. This will be the case, e.g., if you think that genealogical factors (e.g., the identity of *i*'s biological parents) contribute to determining *i*'s race (and are not screened off by

https://doi.org/10.1086/673721 Published online by Cambridge University Press

implication, then why think that it is the right concept for economists or epidemiologists studying race to be using? An argument is needed here to justify the claim that these scientists should work with such a concept of race.

There are thus good reasons to think that the experiment described above is logically, nomologically, and conceptually possible and, so, good reasons to think that it is possible for race to be a treatment in an experiment, even a randomized experiment. It might be, of course, that the relevant notion of possibility is not a logical or conceptual or nomological possibility. And it might be that the concept of race economists and epidemiologists—among others—ought to adopt is in fact one which implies that values of b_i and e_i are not sufficient to determine i's race. It should be clear, however, that one must commit to rather specific views of race and of the notion of possibility at work in HR—and have a good justification for these commitments—in order to defend the view that race violates this rule. In brief, then, even if HR was the right rule for advocates of the CFA to follow, a view I have argued against, it is doubtful that its application would yield the conclusion that race is an attribute or necessary property.

5. A Positive Argument for Race Being a Cause. Consider an imaginary society in which there are two exclusive and exhaustive racial groups, *A* and *B*. Assume that there is a wage gap between *A*s and *B*s in this society: *A*s receive wages that are uniformly 30% lower than the wages received by *B*s 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); for example, they received the same degree from the same school, they have the same work experience, they have the same interpersonal skills, they are equally productive, they have the same preferences regarding wages, and so on. Assume, finally, that there is only one employer in this society and that this employer fixes the wages of every worker.

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 *B*s. 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 *A*s, is the cause of her receiving wages 30% lower than those of w_B. And,

values of b_i). Thus, if you think that races are biological groups unified by genealogical relations (see, e.g., Hardimon 2012), then you should think that what the experiment described above randomly assigns is not genuinely race.

https://doi.org/10.1086/673721 Published online by Cambridge University Press

given the setup described in the previous paragraph, it seems intuitively correct to say that, had w_A been a *B* instead of an *A*, she would have received higher wages.

This commonsensical explanation is a causal explanation since it purports to explain the wage gap by citing its causes, and one of the causes it invokes is the race of w_A and of other *A* workers. This explanation thus is unavailable to those holding the view that race is not a cause. Indeed, according to advocates of the CFA, counterfactuals about what the wages of w_A would have been like had she been a *B* instead of an *A* have impossible antecedents. But what might then explain the wage gap between *A*s and *B*s? I examine the most prominent alternative explanation below.

According to the view defended by Greiner and Rubin (2011), among many others, races themselves play no causal role in generating the wage gap between *A*s and *B*s. What causally explains this gap, rather, are perceptions of race. More precisely, what explains the fact that w_A receives wages 30% lower is not her race in combination with the employer's belief regarding the relative worth of the work of *A*s but the perception of her race by the employer in combination with this same belief. According to this view, then, coefficients attached to variables representing race in models should be understood as representing the causal effect of perceptions of race rather than the causal effect of race itself. There are several problems with this alternative explanation, however. I examine three below.

First, if the move to perceptions is warranted in the case of race, then why should it not be warranted for other properties of units as well? Why not think that, rather than work experience (or education, or . . .), it is the perception of work experience (or education, or . . .) that is causally relevant to an individual's wages, for instance? The move from race to perceptions of race seems rather ad hoc and, in the case of Greiner and Rubin at least, is largely motivated by the assumption that race is not a cause, an assumption that, I argued in section 4, is unjustified.

Second, in the imaginary society I described, it is easy enough to determine who's perception it is that is causally 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 *A*s were on average, rather than uniformly, 30% lower than those of *B*s? Who's perception would then be causally relevant? The collective perception of all the employers? Or the collective perception of only those employers who believe the work of *A*s to be worth less than that of *B*s? 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 in answering these questions as one limitation of this approach (2011, 783–84). 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 inter-

https://doi.org/10.1086/673721 Published online by Cambridge University Press

pretation in terms of perceptions of race of the causal effects estimated by these studies? The move from race to perceptions of race thus raises as many questions as it answers.

Third, what is it that causes the employer in the imaginary society I described to perceive *A* workers (e.g., w_A) to be *A*s? If race is not a cause, then what causes the employer to perceive w_A to be an *A* cannot be the fact that she is an *A*; that is, it cannot be her race. The most plausible alternative here seems to be to claim that what causes the employer to perceive w_A to be an *A* is the instantiation by w_A of a set of features *F* the presence of which is strongly correlated with, but does not constitute, being a *A*. Consider the case in which $F : \{\text{skin color S}\}$. A question immediately arises: If the employer perceives w_A to be an *A* solely on the basis of her skin color and then proceeds to give her wages 30% lower than *B*s in an equivalent job on the basis of this perception, then is this case properly described as a case of racial discrimination? Or is it a case of discrimination on the basis of skin color?

Insofar as the employer de facto equates race and skin color when, by assumption, they are not identical, it seems more appropriate to describe this case as one of discrimination on the basis of skin color than as one of genuinely racial discrimination. Consider the fact that, if the correlation between being an *A* and being of skin color S is less than perfect, then the employer will discriminate against some non-*A*s and fail to discriminate against some *A*s. In other words, the line between workers that are discriminated against and workers that are not will cut across racial groups to follow the line between skin colors. The view that this case is not one of racial discrimination is further supported by standard definitions of 'racial discrimination', for example, the definition formulated by a panel of the US National Research Council that equates racial discrimination with "*differential treatment on the basis of race* that disadvantages a racial group" (Blank, Dabady, and Citro 2004, 39).

In brief, 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.[9] In other words, it is doubtful that Greiner and Rubin can explain cases of genuinely racial discrimination without assuming race to be a cause.

The alternative explanation developed by Greiner and Rubin thus does not seem nearly as satisfactory as the commonsensical explanation sketched above that assumes race to be a cause. Of course, Greiner and Rubin's approach is not the only possible alternative, and it has not been fully worked

9. And so it is not clear that these perceptions are properly called 'perceptions of race' in the first place.

https://doi.org/10.1086/673721 Published online by Cambridge University Press

out yet. But it is by far the most prominent in the literature. That it faces significant difficulties thus provides some support for the claim that one must assume race to be a cause in order to explain racial discrimination.

6. Conclusion. I have defended the view that the argument developed by advocates of the CFA against race being a cause is unsound because two of its premises are false. And I have sketched a positive argument to the effect that race must be assumed to be a cause in order to explain instances of racial discrimination. There thus seem to be good reasons not to follow Holland, Rubin, and other advocates of the CFA in a wholesale dismissal of attempts to draw causal inferences about race as "incoherent" (Greiner and Rubin 2011, 775).

I have said little so far about debates in the philosophy of race. If the arguments developed above are sound, then it seems that philosophers of race should ensure that, whatever concept of race they think ought to be used by scientists studying the role of race, their account of this concept implies that race can be a cause.

The debate over the causal status of race examined in this article also gives a useful example of a case in which 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.

REFERENCES

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Blank, Rebecca, Marilyn Dabady, and Constance Citro, eds. 2004. *Measuring Racial Discrimination*. Washington, DC: National Academies.

Glymour, Clark. 1986. "Statistics and Causal Inference: Comment; Statistics and Metaphysics." *Journal of the American Statistical Association* 81:964–66.

Greiner, James, and Donald Rubin. 2011. "Causal Effects of Perceived Immutable Characteristics." *Review of Economics and Statistics* 93:775–85.

Hardimon, Michael. 2012. "The Idea of a Scientific Concept of Race." *Journal of Philosophical Research* 37:249–82.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–60.

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

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

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

https://doi.org/10.1086/673721 Published online by Cambridge University Press