Causal Agnosticism about Race

Alexander Williams Tolbert[1*]

Abstract

This paper introduces the concept of causal race agnosticism and attempts to provide a detailed examination of the challenges in determining whether race causes specific outcomes. The argument is anchored in two premises: (1) when a causal hypothesis is confounded by numerous unmeasured variables, withholding judgment on the causal relationship may be warranted, and (2) the hypothesis that socially constructed race causes a particular outcome is confounded by many unmeasured variables. The paper explores the challenges posed by violations of two key assumptions in causal inference: positivity and the Stable Unit Treatment Value Assumption (SUTVA). These violations can lead to confounding, complicating causal claims about race. The argument for causal agnosticism highlights the critical importance of addressing confounding and stability of causal relationships in variable selection, which are central to the decision-making process in causal inference research.

1 Introduction

The causal status of race has been a subject of intense debate in the causal modeling and statistics literature [1–4]. This paper contributes to this ongoing discussion by defending the notion of *causal agnosticism* about race, asserting that it is reasonable to withhold judgment on whether race is a cause. The central thesis is that it is reasonable for all socially constructed races to remain agnostic about their causal influence on any outcome. This position constitutes an epistemic claim concerning the reasonableness of causal claims about race, distinct from metaphysical claims about the nature of race itself. The argument for causal agnosticism highlights the critical importance of addressing confounding and stability of causal relationships in variable selection, which are central to the decision-making process in causal inference research.

The significance of this thesis lies in its contribution to the relevant literature on causal modeling, statistics, and the philosophy of race. By addressing the question of race's causal status, this paper engages with fundamental issues in scientific methodology. It provides valuable insights into the complexities of race as a potential causal factor. The subsequent sections of this paper will develop and defend the argument for causal agnosticism about race. In Section 2, I present the argument and clarify critical terms. Section 3 will delve into the defense of the first premise, known as the *causal premise*, which establishes the foundation for the stance of causal agnosticism. Section 4 will delve into the defense of the second premise, known as the *race is confounded*, establishing that socially constructed race has many unmeasured confounders. I examine the existence of confounders in the context of race in two subsections: Subsection 4.1 will analyze the relationship between race and positivity violations.

In contrast, Subsection 4.2 will explore how race can violate the *Stable Unit Treatment Value Assumption (SUTVA)*. In Section 5, I address potential objections and provide robust responses to further strengthen the case for causal agnosticism. Finally, in Section 6, I offer concluding remarks and propose future directions for research in understanding the complex interplay between race and causality.

2 Argument for Causal Agnosticism about Race

My central argument for *causal race agnosticism* rests on two key premises:

> **Premise 1 (Causal Premise):** For all $\mathcal{A}$ and $\mathcal{Y}$, if the hypothesis that $\mathcal{A}$ causes $\mathcal{Y}$ has many unmeasured confounders, then it is reasonable to withhold judgment as to whether $\mathcal{A}$ causes $\mathcal{Y}$.
>
> **Premise 2 (Race is Confounded):** For all R and $\mathcal{Y}$, if R represents a socially constructed race, then the hypothesis that R causes $\mathcal{Y}$ has many unmeasured confounders.
>
> **Thesis (Causal Agnosticism about Race):** For all R and $\mathcal{Y}$, if R represents a socially constructed race, it is reasonable to withhold judgment whether R causes $\mathcal{Y}$.

This section lays the foundation for my argument for causal agnosticism about race by clarifying critical terms in the premises. Social constructionist views of race posit that race is a non-biological but real social kind, shaped by historical, political, and economic factors rather than innate biological traits [5]. Next, I define “unmeasured confounding” in the causal modeling literature. According to Ananth and Schisterman [6, p. 1], an unmeasured confounder is a variable connected to independent and dependent variables and could account for the observed relationship. Failure to account for all relevant confounding variables can lead to biased treatment effect estimates and invalidate causal inferences.

Additionally, I explore the fundamental assumptions of causal inference, namely the *positivity* and *Stable Unit Treatment Value Assumption (SUTVA)*. Positivity assumes a positive probability of receiving the treatment for all units in the study population, given their observed covariates [7, p. 30-31]. A positivity violation occurs when some units have a zero probability of receiving the treatment, given their covariates. SUTVA assumes that the potential outcome of any unit is not affected by the treatment assignment of other units and that only one version of the treatment is being studied [7, p. 5-6]. A violation of SUTVA occurs when this assumption is unmet, making it difficult to disentangle the actual treatment effect. These assumptions are crucial in ensuring the validity of causal inferences. Finally, I address the context-dependent nature of the threshold for withholding judgment when dealing with "many" unmeasured confounders. While a single strong unmeasured confounder may cast doubt on a causal claim, many weakly related unmeasured confounders might not be sufficient to justify withholding judgment. Evaluating the strength and relevance of each potential unmeasured confounder is crucial to making informed decisions about causal claims.

Having established the concepts of the social construction of race, unmeasured confounding, and assumptions in causal inference, I now explore how these ideas intersect when studying race and its causal influence. In the following section, we will present the Causal Premise and build upon these concepts to establish the core argument for causal agnosticism about race.

3 Defending the Causal Premise

The causal premise is well established in the causal modeling literature as being required mathematically for establishing an unbiased estimate of the effects of surgical intervention [7]. Intuitively the *causal premise* discusses the relationship between two variables, which we will call $\mathcal{A}$ and $\mathcal{Y}$. $\mathcal{A}$ represents an "action" or intervention and $\mathcal{Y}$ represent some outcome of interests. To say that $\mathcal{A}$ causes $\mathcal{Y}$, we need to rule out the possibility that some other factor, such as $\mathcal{X}$, is responsible for our observed relationship. One way to do this is to measure all the relevant variables and control for them statistically. However, if many unmeasured variables could be confounding the relationship, then we cannot be sure that we have ruled out the possibility of alternative explanations. In this case, the causal premise says that it is reasonable to withhold judgment about whether $\mathcal{A}$ causes $\mathcal{Y}$. Dawid highlights a critical consideration in his work [8, p.282-283]: the question of whether the observed distribution of outcomes for individuals who underwent an intervention can serve as a stand-in for the hypothetical distribution of outcomes had they not undergone the intervention. Here, the term 'distribution of outcomes' refers to the array of potential outcomes an individual might experience following an intervention.

When we use the observed outcome distribution of individuals who received an intervention to approximate their hypothetical distribution without the intervention, we actively assume that their observed distribution under the intervention reasonably approximates what we would have observed if they had not undergone the intervention. In simpler terms, we use the outcome distribution of individuals who received the intervention to estimate the distribution of outcomes had the intervention not been given. We assume that individuals who underwent the intervention are significantly comparable to those who did not. We attribute any differences between these groups to other factors, such as demographic characteristics or health status. To achieve this, we assume what is commonly called 'exchangeability' between the intervention recipients and ourselves concerning relevant pre-intervention features.

However, as Dawid [8] points out, this principle of exchangeability might not apply to the subset of individuals who underwent treatment, as these characteristics could have influenced the decision to administer treatment. For example, a particular medication may have been prescribed only to individuals of a certain race at higher risk for a specific condition. This association may induce a spurious or distorted relationship between the exposure and outcome, such that the actual causal effect is obscured. This scenario is commonly known as *confounding*, which hinders interpreting causal relationships in observational data.

Achieving exchangeability can be done through randomization. Randomized controlled trials (RCTs) are usually considered the standard for determining causal impacts. However, they may not always be possible, especially if the variable under consideration is non-manipulable, such as "race." In such situations, it may be necessary only to make conclusions once technological advancements allow for manipulating the treatment variable.

Consider a setting where uppercase variables are random variables and lowercase variables are their realizations. The population average causal effect (ACE) of an intervention on an outcome can be estimated by considering the data domain $\mathcal{D} = \mathcal{Y} \times \mathcal{X} \times \mathcal{A}$. The average causal effect (ACE) compares the hypothetical intervention of everyone receiving the treatment with everyone not receiving it for a well-ordered target population. The difference between these two scenarios is the expected value of the potential outcome under intervention, $\mathrm{E}[\mathcal{Y}(1)] - \mathrm{E}[\mathcal{Y}(0)]$, where $\mathcal{Y}(a)$ is the potential outcome under intervention a. We rely on certain assumptions to estimate the average treatment effect from observational data. The following sections will focus on two assumptions, positivity and SUTVA, and their role in establishing ACE.

4 Defending the Race is Confounded Premise

4.1 Race and Positivity

Recall that failures of positivity are one of two reasons supporting my agnosticism. Positivity is a crucial assumption in the causal inference that ensures that all population subgroups have some chance of receiving the treatment. This assumption is essential for estimating the treatment effect in those subgroups. *Deterministic* and *stochastic* positivity are two types of positivity confounding that can affect the validity of causal inference [9]. To achieve deterministic positivity, every group within a population must have a chance of receiving treatment. This assumption requires that if there is a treatment variable called $\mathcal{A}$ and a set of covariates known as $\mathcal{X}$, there should be a positive probability of receiving both levels of treatment $\mathcal{A}$ for every value of the covariates $\mathcal{X}$.

Violating deterministic positivity can lead to bias by providing inaccurate estimates of the treatment effect for subgroups not exposed to it. In a dataset, specific subgroups may have zero probability of exposure to a particular intervention. Deterministic positivity is violated when the studied variables require the absence of treatment. For example, it is impossible to include an 80-year-old pregnant woman or a 5-year-old with a Ph.D. in a study of pregnant women.[1] Similarly, if a study examines the effects of a new medication on heart disease and the medication is only given to patients who have already had a heart attack. Deterministic positivity is violated because patients without a heart attack cannot receive the drug.

[1] Thanks to Jay Kaufman for providing this creative example over email correspondence

Stochastic positivity, on the other hand, requires that every possible sample of individuals from the population has some chance of receiving both levels of the treatment. The formal definition of stochastic positivity states that for a treatment variable $\mathcal{A}$, a set of covariates $\mathcal{X}$, and a sample size n, stochastic positivity occurs if there is a positive probability of receiving both levels of the treatment $\mathcal{A}$ for every value of the covariates $\mathcal{X}$ and every sample size n. This ensures that the treatment effect can be estimated in all possible instances, regardless of their characteristics. Stochastic non-positivity is a finite sampling issue due to the inherent variability in data collection.

Violations of stochastic positivity can lead to bias because the treatment effect cannot be evaluated in samples that are not representative of the population. For example, a study investigating the effect of a new therapy on depression violates stochastic positivity if the sample only includes individuals who are already receiving treatment for depression, as individuals who are not receiving treatment for depression have no chance of being included in the selection. Deterministic positivity and stochastic positivity are crucial concepts that ensure the validity of causal inference.

Now with the postivity and its variations explained, the rest of this section will show that race can violate both the stronger definition of deterministic postivity and the weaker definition of stochastic positivity, which leads to confounding. Remember, confounding arises when a variable is linked to both the predictor and the outcome, making it difficult to ascertain the causal relationship between the predictor and the outcome. In the social sciences, several studies have remarked on the confounding of race with socioeconomic status [10]. For example, consider race and socioeconomic status (SES). The work of Messer et al. [11], as highlighted by VanderWeele and Robinson [3, p.477], underscores the complexities inherent in differentiating the effects of SES and race. They point out that in situations characterized by significant income disparities, where individuals from a specific racial group exclusively occupy a particular SES, it becomes exceedingly difficult to disentangle the influences of SES and race. This difficulty persists even when data on these variables are available. The absence of overlap between race and SES complicates isolating each variable's causal influence. [12] show that in the context of SES and race, these violations are primarily associated with the structural patterns created by segregation and racism. Segregation and racism structure the data in a way that highly correlates with race and poverty in specific areas of the United States. However, it is essential to note that these patterns are not deterministic, meaning there are exceptions to the general trend. For example, there are affluent Black neighborhoods and poor White neighborhoods, although these instances may be relatively sparse in specific geographic settings.

Consider another example of positivity violations with race. Suppose we want to study the effect of race on the risk of developing skin cancer. We have data on race (black vs. white) and skin pigmentation, which we measure on a continuous scale ranging from 0 (lightest) to 1 (darkest). We also have data on other potential confounders like age, sex, and sun exposure. We hypothesize that darker skin pigmentation is protective against skin cancer, and we want to estimate the causal effect of race on skin cancer risk while adjusting for skin pigmentation and other potential confounders. In this case, the positivity assumption requires that some individuals from both racial groups exist within each level of the covariates. We should observe some black and some white individuals for any combination of covariate values. However, in our example, we may not have any dark-skinned white individuals in our sample. This observation means that we would not be able to estimate the effect of race on individuals with dark skin pigmentation. [2] Similarly, we may not have any light-skinned black individuals in our sample, which would prevent us from estimating the effect of race on individuals with light-skin pigmentation. This result is an example of how skin pigmentation can violate the positivity assumption, making it difficult or impossible to estimate the causal effect of race on a health outcome.

The absence of individuals from a marginalized group in privileged societal positions leads to substantial positivity violations, some of which may even be deterministic. This situation arises when discriminatory practices and social structures systematically prevent marginalized group members from attaining positions of power, privilege, or access to resources. For example, historical instances like the caste system in India, the antebellum South in the United States, or the occupation of Ireland by the British exemplify extreme discrimination that resulted in deterministic positivity violations. In these cases, the discriminatory systems were deeply entrenched, creating structural barriers that limited the opportunities for individuals from marginalized groups to advance socioeconomically or gain access to higher social positions. In India's caste system, a hierarchical social structure based on birth and occupation, individuals were assigned to specific castes, and mobility between castes was extremely limited. The discrimination embedded within this system ensured that individuals from lower castes faced significant barriers to upward mobility, resulting in a lack of overlap between the higher castes and the discriminated lower castes. This deterministic positivity violation results from systematically denying privileged positions and opportunities to individuals from marginalized castes.

Similarly, during the antebellum period in the Southern United States, slavery and racial segregation were widespread. Enslaved and marginalized African Americans were denied education, economic opportunities, and political power. The systemic discrimination and the institution of slavery ensured that individuals from the enslaved population could not occupy privileged positions in society. As a result, there was a lack of overlap between African Americans and privileged White individuals, representing a deterministic positivity violation in racial discrimination. The occupation of Ireland by the British provides another example. The oppressive policies and discriminatory practices imposed by the British colonial rule limited the opportunities for the Irish population to gain social and economic advantages. The systematic denial of privileges and resources to the Irish population resulted in a deterministic positivity violation, where individuals from the occupied Irish community were effectively excluded from positions of power and influence held by the British occupiers.

These historical examples illustrate how extreme discrimination can create deterministic positivity violations by structurally limiting the overlap between privileged and marginalized groups. The systemic denial of opportunities, resources, and privileges to individuals from the discriminated group perpetuates a state where the marginalized group remains excluded from privileged positions. This, in turn, contributes to the absence of certain combinations or subgroups in the data, leading to

[2] Thanks to Zack Lipton for this creative example

positivity violations. Understanding these dynamics of extreme discrimination and their impact on positivity violations is crucial for recognizing the structural barriers within societies and informing efforts toward equity, justice, and dismantling discriminatory systems.

Economists Hamilton and Darity Jr [13] argue that the broken promise of 40 acres and a mule to ex-slaves, coupled with systematic property deprivation of Black Americans between 1880 and 1910, has led to a racial wealth gap in the U.S. This gap is perpetuated by structural barriers fueled by past and present racial exploitation and discrimination. Inheritances, bequests, and intra-family transfers contribute more to this gap than education and income, with white families receiving larger estates on average than African American families. More generally, anytime you have the intense stratification of social groups, as with racial wealth, a lack of overlap hinders causal estimation epistemically.

Further, metaphysical consequences and conceptual dilemmas arise for some views of race in the philosophy of race literature. For example, Sally Haslanger's social construction view of race is defined as racialized practices constructing social realities around physical features. In their account, "hierarchical positioning of an ethnic group within a broader society (or broader political formation) is a process of racializing the group" [5, p. 27]. I refer to this as Haslanger's *racialization* thesis. Haslanger claims that racialization is the process by which social races are formed [5].

The racialization thesis creates odd tensions in knowing the causal effects of *race*. Tension arises when a group is increasingly subjected to racialization, leading to heightened discrimination against them. As a result, there will not be any members of that group in privileged positions in society, creating a lack of overlap between the groups and a violation of positivity. A dilemma has now become apparent. If Hasslangerian social constructivists hold that racialization is accurate, the more they believe it holds, the less they can be sure about race's causal effects because of positivity violations. However, if they deny the force of racialization in our current society to know race's effect, they deny the existence of race in Hasslanger's terms. They must trade knowledge claims about race's causal powers with metaphysical claims about race's existence as social construction according to how they conceive it. One ethical implication from my argument is that to solve our epistemic problem; we may need to solve our moral one first. Hamilton and Darity Jr [13], mentioned earlier, focused on normative grounds for reparations for reasons rooted in the racial wealth gap. However, my arguments suggest that we also have epistemic reasons that can serve as grounds for repair. Fixing the social-political problem of social inequality among groups contributing to the lack of postivity would make causal inference easier. This epistemic conundrum may open another route for generating normative obligation for reparations and other forms of social egalitarianism.

4.2 Race and Violations of SUTVA

Recall that the failures of SUTVA are the last of my two reasons supporting my agnosticism. In causal inference, the selection of variables is of utmost importance. As highlighted by Woodward [14], stability is a crucial consideration that should constrain variable choice. Stability refers to the degree to which a causal relationship between two variables, $\mathcal{X}$ and $\mathcal{Y}$, is generalizable from one set of circumstances to another. The concept of stability acknowledges that a causal relationship between $\mathcal{X}$ and $\mathcal{Y}$ may hold under certain conditions but not in others. More superficially I take stability, to refer to the uniformity of a conditional distribution across different realizations of the data. This means that the conditional distribution of Y given X and the non-stochastic regime F_X which indexes the different conditions under which a system is observed remains the same regardless of the specific conditions or state under which the data is generated.[3] Mathematically, this can be represented as $P(Y|X, F_X)$ being invariant under changes in the realization of X, Y, and F_X. Stability ensures that the probabilistic relationships are the same across different regimes. Consequently, selecting variables that allow for the characterization of more stable causal relationships is essential.

In science, unstable variables have been a persistent problem. As noted by Woodard, many of these variables, including total cholesterol, lack stability due to the possibility of ambiguous manipulations. This ambiguity is further explained by Spirtes and Scheines [17], who notes that stability refers to the ability to make unambiguous manipulations of a variable. Scholars such as Spirtes and Scheines [17] and Chalupka et al. [18] have studied the effect of cholesterol on heart disease by examining the aggregate of low-density (LDL) and high-density (HDL) cholesterol. They found that LDL and HDL exert completely distinct impacts on heart disease. This variability leads to an ambiguous association between total cholesterol and heart disease. In certain situations, higher levels of total cholesterol increase the risk of heart disease.

Conversely, these higher levels may reduce the risk of heart disease in other situations. Hence, the utility of total cholesterol as a causal variable becomes limited. It encapsulates the combined measure of LDL and HDL, each exerting distinct effects on heart disease. For instance, prescribing a “low-cholesterol diet” entails an ambiguous intervention. As noted by Chalupka [19, p.9], the term ”low cholesterol” could imply a reduction in LDL, HDL, or both, with each possibility having distinct implications for cardiovascular health. Without prior knowledge about the relative levels of LDL and HDL, it becomes challenging to conclusively determine the cause-and-effect relationship between total cholesterol levels and the incidence of heart disease. Woodward [14] claims that in such situations, microvariables like LDL and HDL emerge as preferable alternatives since they exhibit more consistent effects on heart disease.

Similar problems arise with the cluster-kind views of race. By cluster-kind views of race, I mean a theory that posits race as a higher-order macrovariable constitutive of lower-order microvariables. Social science has cluster-kind accounts of race that posit race is constitutive of many variables. For instance, Sen and Wasow [20, p.506] suggest a variety of potential variables that could be considered, including ancestry region, wealth, dialect, genetic factors, neighborhood characteristics, diet, social standing, norms, power dynamics, class, skin color, religion, and region of origin. On the other hand, VanderWeele and Robinson [3, p.7] propose a more streamlined cluster view of race that includes fewer potential variables such as physical attributes, parental physical attributes, genetic heritage, cultural environment, family socioeconomic status,

[3] For more technical details on regime indicators see [8, 15, 16]

and neighborhood socioeconomic status. These perspectives portray race as a complex and multifaceted concept. Various "cluster-kind views of race" propose different variables that could contribute to constructing the concept of *race.*

Now, let us consider the potential violation of SUTVA. Suppose researchers are conducting a study to investigate the relationship between race and the effectiveness of a new drug in treating a specific health condition. They consider two approaches to recruiting participants based on race: one focusing on genetic ancestry and another on self-identification or other factors. In the first approach, researchers aim to recruit participants with similar genetic backgrounds within racial groups. They believe that by creating a more uniform pool based on ancestry, they can isolate the effects of race on the outcome variables. However, this approach introduces a potential challenge related to the drug's effectiveness across different genetic backgrounds. If the drug works well for specific genetic backgrounds within a racial group but not for others, selecting a more uniform pool based on ancestry could mask these differences. The researchers might observe inconsistent results because the effects of the drug might vary among individuals with different genetic backgrounds, even within the same racial group.

Consequently, drawing valid causal inferences about the drug's effectiveness becomes challenging as the relationship between race and the outcome variables becomes unstable. On the other hand, the second approach involves recruiting participants based on self-identification or other factors associated with race. This approach acknowledges the complexity and multifaceted nature of race, including social and cultural factors, and aims to capture the lived experiences of individuals within different racial groups. However, this approach also introduces potential confounding variables that can affect the causal inferences regarding the drug's effectiveness. Researchers might inadvertently introduce confounding variables associated with race and outcome variables by relying on self-identification or other factors to recruit participants based on race. These confounding variables can complicate the analysis and interpretation of the causal relationship between race and the drug's effectiveness, as it becomes difficult to disentangle the specific effects of race from other factors that might influence the outcomes.

In social scientific models, race is often used as a single variable, despite the substantial heterogeneity within racial categories. Such an approach is akin to using total cholesterol as a predictor in heart disease models. Total cholesterol combines LDL and HDL, each with differing impacts on heart health, thus, masking their unique effects. A salient example is using the "Black" category in income disparity analyses. The "Black" category aggregates individuals from a wide range of ethnic and ancestral backgrounds, like those tracing their roots to various regions of Africa, Afro-Caribbean individuals, or those residing in the United States for multiple generations. Using "Black" as an aggregate variable combines diverse experiences, backgrounds, and social factors within this category. Consider the contrasting experiences of a recent Nigerian immigrant and a U.S. resident descended from enslaved Africans, which underscore the complexities within the "Black" category. On the one hand, recent Nigerian immigrants, potentially benefiting from selective immigration policies, may attain higher levels of education, enabling access to lucrative professional opportunities and potentially resulting in higher income [21].

On the other hand, descendants of enslaved Africans in the U.S. grapple with a markedly different socioeconomic landscape. As mentioned earlier enduring legacy of slavery, unfulfilled restitution promises, and systemic property deprivation have shaped their economic reality, perpetuating a persistent racial wealth gap in the U.S. A significant facet of this wealth gap is the intergenerational transfer of wealth [13]. The systemic discrimination and socioeconomic barriers African Americans face often limit their ability to accumulate wealth, curtailing the value of inheritances and intra-family transfers. These factors collectively relegate many within this subgroup to lower income strata, highlighting the heterogeneity often obscured within the 'Black' category making the causal relationships with income ambiguous.

In this context, disaggregated microvariables, such as specific ancestry, immigration history, and the degree of systemic racial discrimination, even wealth may offer more stable and thus better income predictors than the macrovariable of race. These variables might present more stable effects on income, enabling a clearer understanding of the causal relationships involved. Consequently, the race variable's stability and its causal relationship with outcomes could be questioned.

These intricacies, along with the violations of positivity and the Stable Unit Treatment Value Assumption (SUTVA), introduce multiple unmeasured confounders for race, prompting us to withhold judgment about its causal effects. Ensuring stable causal relationships and understanding the appropriate level of aggregation for variables is essential for deriving universally applicable insights, which are particularly valuable for devising effective interventions and controls.

5 Objections and Replies

5.1 Objection

This section addresses an objection to the claim that race violates the positivity assumption. The objection argues that certain variables, such as SES or neighborhood, should not be included in the conditioning set because they are post-exposure variables that are descendants of racialized status. This objection asserts that including post-exposure variables in the conditioning set would block essential pathways through which race affects outcomes and would not violate positivity. A post-exposure variable is a variable that is measured or observed after an individual has been exposed to a particular intervention of interest and is often used in causal inference and modeling to evaluate the effects of the exposure or treatment on a particular outcome of interest. Recall positivity assumption is formally stated as $P(\mathcal{A} = a|X) > 0$, where $\mathcal{X}$ is the set of variables that are sufficient to adjust for in identifying the ACE, the same X that appears in $\mathcal{Y}(a) \perp\!\!\!\perp \mathcal{A}|\mathcal{X}$. These $\mathcal{X}$s are confounders, not mediators or anything post-exposure (descendants of $\mathcal{A}$ would not be in $\mathcal{X}$).

The variables that should be included in $\mathcal{X}$ depend on what might confound the $\mathcal{A}$, $\mathcal{Y}$ relationship, and factors influencing racialized group membership and the outcome. For example, age and gender might be included because age is correlated with race by various mechanisms, and gender plausibly affects racialization according to intersectionality theory. However, $P(\mathcal{A} = a|\text{age, gender}) > 0$ should not pose a problem in any reasonable study design. However, variables like SES or neighborhood are clearly "post-exposure" because they are causally descendants of racialized status, not the reverse. The objection claims that a good study design would never include them in $\mathcal{X}$. [4]

5.2 Reply

Although SES and wealth are typically considered post-exposure variables downstream of racialized status, the racialization thesis posits that these variables also play a role in the formation and maintenance of race. Therefore, while it is true that controlling for these variables may block one pathway through which race affects outcomes, it is also essential to consider how these variables contribute to the broader social context in which race operates and is formed. In essence, the oppression variable is a causal variable that has a direct causal arrow going into race. By disaggregating the oppression variable, we can identify the variables that have a causal flow into race, such as SES, wealth, neighborhood, and geography. The objection to including certain variables, such as wealth and SES, in the adjustment set $\mathcal{X}$ for a given analysis is that these variables may not be necessary confounders. However, this objection needs to recognize that these variables are not just downstream effects of race but also contribute to the formation and maintenance of race itself. In other words, the causal relationship is cyclic, and these variables have a causal arrow going directly into race.

For example, imagine we want to study the effect of race on academic achievement in a particular school district. We might suspect that SES is a confounding variable, as it is known to be associated with race and academic achievement. If we adopt SES as a post-exposure variable that should not be included in the adjustment set $\mathcal{X}$, we would only adjust for variables confounders of the race-academic achievement relationship and not descendants of race. For example, we might adjust for variables such as parental education level, number of siblings, and distance from school. However, if we adopt the position that racialization theory holds, we would include SES in the adjustment set $\mathcal{X}$. We would include SES because racialization theory posits that oppression and SES are among the factors that create and maintain race and, thus, are not just descendants of race but also its causes. Therefore, adjusting for SES would be necessary to fully account for the confounding effect of race on academic achievement. It is necessary to balance the need to control for confounding variables with the need to understand the broader social context in which race operates and is formed. This argument underscores the importance of carefully selecting variables for inclusion in causal models and the need to consider what variables may be confounders and which may not be confounders for specific outcomes.

6 Conclusion

In conclusion, the argument for causal race agnosticism is grounded in two essential premises: the *causal premise* and *race is confounded*. The *causal premise* posits that when many unmeasured factors confound the causation hypothesis between two

[4] Thanks to Daniel Malinsky for raising this insightful objection

variables, it is reasonable to withhold judgment about their causal relationship. The premise that *race is confounded* recognizes the complexity of race as a socially constructed concept influenced by various factors. Positivity violations and failures of SUTVA arise as critical obstacles to drawing inference causal conclusions about race's effects. Given these complexities, embracing causal agnosticism about race is a modest epistemic approach. It acknowledges the epistemic limitations due to positivity violations and SUTVA violations in studying race but, more importantly, calls for addressing underlying social and structural issues perpetuating racial inequalities is crucial to reducing confounding and improving causal inference.

References

[1] Holland, P.W.: Statistics and causal inference. Journal of the American statistical Association **81**(396), 945–960 (1986)

[2] Glymour, C.: Comment: Statistics and metaphysics. Journal of the American Statistical Association **81**(396), 964–966 (1986)

[3] VanderWeele, T.J., Robinson, W.R.: On causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology (Cambridge, Mass.) **25**(4), 473 (2014)

[4] Glymour, C., Glymour, M.R.: Commentary: race and sex are causes. Epidemiology **25**(4), 488–490 (2014)

[5] Glasgow, J., Haslanger, S., Jeffers, C., Spencer, Q.: What Is Race?: Four Philosophical Views. Oxford University Press, ??? (2019)

[6] Ananth, C., Schisterman, E.: Hidden biases in observational epidemiology: the case of unmeasured confounding. BJOG: an international journal of obstetrics and gynaecology **125**(6), 644 (2018)

[7] Hernan, M., Robins, J.: Causal Inference: What If. Boca Raton: Chapman & Hill/crc, (2020)

[8] Dawid, A.P.: Statistical causality from a decision-theoretic perspective. Annual Review of Statistics and Its Application **2**, 273–303 (2015)

[9] Zivich, P.N., Cole, S.R., Westreich, D.: Positivity: Identifiability and estimability. arXiv preprint arXiv:2207.05010 (2022)

[10] LaVeist, T.A., Thorpe Jr, R.J., Mance, G.A., Jackson, J.: Overcoming confounding of race with socio-economic status and segregation to explore race disparities in smoking. Addiction **102**, 65–70 (2007)

[11] Messer, L.C., Oakes, J.M., Mason, S.: Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding.

American journal of epidemiology **171**(6), 664–673 (2010)

[12] LaVeist, T.A.: Segregation, poverty, and empowerment: health consequences for african americans. The Milbank Quarterly, 41–64 (1993)

[13] Hamilton, D., Darity Jr, W.: Can 'baby bonds' eliminate the racial wealth gap in putative post-racial america? The Review of Black Political Economy **37**(3-4), 207–216 (2010)

[14] Woodward, J.: The problem of variable choice. Synthese **193**(4), 1047–1072 (2016)

[15] Dawid, A.P.: Causal inference without counterfactuals. Journal of the American statistical Association **95**(450), 407–424 (2000)

[16] Dawid, P.: Decision-theoretic foundations for statistical causality. Journal of Causal Inference **9**(1), 39–77 (2021)

[17] Spirtes, P., Scheines, R.: Causal inference of ambiguous manipulations. Philosophy of Science **71**(5), 833–845 (2004)

[18] Chalupka, K., Eberhardt, F., Perona, P.: Causal feature learning: an overview. Behaviormetrika **44**(1), 137–164 (2017)

[19] Chalupka, K.: Automated macro-scale causal hypothesis formation based on micro-scale observation. PhD thesis, California Institute of Technology (2017)

[20] Sen, M., Wasow, O.: Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. Annual Review of Political Science **19**(1), 499–522 (2016)

[21] Sakamoto, A., Amaral, E.F., Wang, S.X., Nelson, C.: The socioeconomic attainments of second-generation nigerian and other black americans: Evidence from the current population survey, 2009 to 2019. Socius **7**, 23780231211001971 (2021)