Legal Theory Workshop
UCLA School of Law

Lily Hu

Yale University

"Discrimination, Causation, and
Detecting Normative Phenomena"

Thursday, March 19, 2026, 3:20pm – 5:20pm
Law School Room 1314

Draft, February 2026. For UCLA Workshop.
Please don't cite or quote without permission.

Note to UCLA Legal Theory workshop participants:

This is *a very rough draft*. I even hesitate to call it a "draft" because that label seems to commit to future iterations of the ideas here, which I am very much doubtful of. But I thought I'd present something which would very much benefit from being "workshopped."

This topic was originally envisioned as something I'd like to include a chapter about in my book project on causal analyses of race. Most of the book is about the methodological foundations of causal inference about race and the tricky problems that come up when we think about race as a cause within the leading interventionist framework of causation. I had the thought to include a chapter on the very popular approach of theorizing racial discrimination by way of analyzing race causation and identifying race's causal effects with discriminatory effects. What you are about to read is a first-pass set of thoughts I have about the topic. Where I refer to arguments earlier in the book, I try to offer a brief summary of them, though this certainly does not mean they are therefore clear. I appreciate all your being game for my sharing something that should probably have spent more time in the oven, before others got involved. But together, we might yet be able to turn lead into gold!

Discrimination, Causation, and Detecting Normative Phenomena
Lily Hu

§1. Introduction

In 2001, the Committee on National Statistics convened an interdisciplinary group of experts to study race-based discrimination in the United States. The meeting's opening premise was that contemporary American society is riven with racial inequality and that much of this inequality is the product of ongoing racial discrimination, even decades after the end to its legalized forms. The problem was that the precise relationship between racial disparities and racial discrimination was murky. Answers to such questions as, how to identify discrimination, how much of existing disparities were due to it, and by what avenues it was perpetrated, remained unclear, despite a considerable body of social-scientific research on the topic. The goal of the assembled Panel on Methods for Assessing Discrimination was not just to record the Panel's own considered answers to these questions; even more importantly, it sought to lay the groundwork for a coordinated social-scientific research agenda for the study of discrimination and its effects for the future. "Just because it is challenging to measure discrimination," its report, titled "Measuring Racial Discrimination," said, "does not mean that sound, adequate research in this area is not possible. To the contrary, existing methods and data have produced useful results on particular types of discrimination in particular aspects of a domain or process" (12).

"Measuring Racial Discrimination," is, by all accounts, a work of conventional empirical social science, and leaves to the side any characterization of racial discrimination as a normative phenomenon. The report does not consider questions such as what makes discrimination wrong, what values underlie protections against discrimination, and what exactly constitutes a "protection" against discrimination. It does not ask what animates antidiscrimination laws and norms; nor does it consider what normative conflicts they might produce and how we ought to navigate those conflicts. In fact, the report explicitly avoids substantive ethical theorizing about any issues relating to race and racial inequality. Indeed, the Panel was comprised of a "diverse group of experts in the fields of criminal justice, law, economics, psychology, public policy, sociology, and statistics" (xii). None of its members specialized on the moral and political character of racial discrimination and associated phenomena.

To focus on discrimination as a social and empirical phenomenon is all well and good, especially since the Panel's stated goal was "to define discrimination and review methods for measuring it" (xii). The question arises, however, whether this empirical study of racial

discrimination is possible without an explicitly ethical study of race and race-related harms and wrongs. Is it really possible

> to identify where racial discrimination occurs and to measure the extent to which discrimination may contribute to racial and ethnic disparities, … to consider the [purely empirical] definition of racial discrimination, assess current methodologies for measuring it, identify new approaches, and make recommendations about the best broad methodological approaches,… Give the policy and scholarly communities new tools for assessing the extent to which discrimination continues to undermine the achievement of equal opportunity by suggesting additional means for measuring discrimination that can be applied not only to the racial question but in other important social arenas as well. Conduct a thorough evaluation of current methodologies for measuring discrimination in a wide range of circumstances where it may occur (1–2).

without submitting any moral and political characterization of racial discrimination and nearby race-related normative phenomena?

<p style="text-align:center">***</p>

The implicit answer from "Measuring Racial Discrimination" and the large and ever-expanding body of work on discrimination from the social sciences is a resounding, yes. So, economists, sociologists, political scientists, have all (seemingly) produced reams of studies, analyses, experiments, and methodologies for doing so. And increasingly, the law itself seems to agree: that it is indeed possible to detect racial discrimination with strictly empirical methods that do not advert to any ethical views as to, among other matters, what wrongs or harms are at the heart of the phenomenon or what the purpose of antidiscrimination laws are. So, across numerous discrimination trials, expert social-scientific testimony has (seemingly) brought their specialized techniques for detecting discrimination to bear on the legal claim. The presumption is that it is possible to *define*, *identify*, *measure* discrimination on the basis of race without answering any of the substantive normative questions mentioned above.

The reason for this consensus view is that discrimination on the basis of race, so many seem to agree, is an essentially *causal* phenomenon. Specifically, it occurs when race, in one way or another, causes some adverse outcome. This characterization of racial discrimination as an essentially causal process makes it formally just like so many familiar objects of scientific study and, accordingly, amenable to the standard methods and techniques of empirical science. What distinguishes an investigation into whether discrimination occurred from more conventional

scientific causal investigations into whether, for example, some incidence of lung cancer was caused by smoking or what the effects of drought are on crop failure? If racial discrimination is some species of race causation, then to *detect*, *identify*, and *measure* racial discrimination would seem to consist in the purely descriptive tasks of tracing and delineating race's effects. Normative theorizing may yet be crucial for understanding the broader phenomenon of discrimination, but it may be separated from the purely causal analysis which is required to grasp its empirical dimensions.

This chapter challenges this widespread characterization of racial discrimination as a non-normatively-defined kind of race causation, which I will call the "non-normative race causation" theory of racial discrimination, and in so doing, challenges the widespread presumption that we can detect, identify, and measure racial discrimination without appeal to substantive moral and political judgments about the nature of race, racial inequality, and racial discrimination. This picture is exemplified in "Measuring Racial Discrimination," which is mostly concerned with various race-centered causal methods for assessing discrimination while remaining (supposedly) agnostic about the moral and political values that animate antidiscrimination norms. It is also expressed in the law's willingness to grant causal experts a major role in adjudicating discrimination claims, and in the vast body of social-scientific scholarship wherein authors strive to gain normatively-neutral traction on racial discrimination by analyzing, from an ethically-neutral stance, what race causes.

The view is already considerably undercut by one of the central arguments developed over the course of the first part of the book: that an interventionist causal analysis of race is necessarily normative in the sense of being value-laden. Parts of that argument will thus reappear in this chapter in the course of my showing how the prevailing approaches to analyzing racial discrimination by way of race causation fail to be value-neutral in the way that proponents of the idea intend. But the failure to match this self-conception has wider ramifications. If no causal analysis of racial discrimination can get off the ground without starting with a substantive normative characterization of race, racial inequality, and racial discrimination, as I will argue, the broader projects within which the causal analyses of race are situated must be reconceived and reconfigured. The challenge bears not just on the dominant intellectual and methodological approaches to theorizing racial discrimination but on the wider social practices and structures through which we collectively come to understand and decide what racial discrimination is.

I have already said that the non-normative race causation account is the leading social-scientific view of racial discrimination and is explicitly endorsed by many scholars who work on the topic. But the view is also influential although more implicit in philosophy, and so I will begin in §2

by characterizing it more fully and noting how it appears in both the social-scientific and philosophical literatures on discrimination. As I will explain, this picture of racial discrimination is what justifies the common two-stage approach to analyzing discrimination on the basis of race, which starts first by sorting out the non-normative facts about what race causes, before then moving on to the normative ones.

From here, I proceed in §3 to present an argument against the non-normative race causation approach to racial discrimination, which I take to be defined by its commitment to what I will call the *separability* thesis: the scientific tasks of detecting and measuring discrimination and the normative tasks of morally and politically analyzing discrimination are separable. They can (and should) be carried out in isolation from one another.[1] Having concluded that the standard non-normative causal analysis cannot furnish any theory of discrimination whatsoever, I turn in §4 to draw out the normative stakes of the impossibility of separability. As we will see, there are not just first-order normative concerns about the concept of racial discrimination that is produced from standard value-laden causal inference analyses of race, but moreover second-order concerns, as I will show, about the shape that the broader conversation surrounding discrimination takes—and cannot take—when causal analyses of race occupy a place of priority.

I close in §5…

§2. The "non-normative race causation" theory of racial discrimination

It is perfectly natural to think of discrimination in causal terms. We often talk of discrimination as happening "because of" some feature, pointing to a causal link between the feature and the discriminatory action. Racial discrimination is discrimination "because of" race, which in turn is commonly understood as an adverse action *caused by* race. If I racially discriminate against you, my treatment of you depends on your race. Your race causes my specific action.

This rather explicit framing of discrimination in causal terms does not appear so much in philosophical characterizations of racial discrimination, but I do not take this to reflect any dissensus or uncertainty more generally within the field about understanding discrimination as a causal notion.

---

[1] On top of the separability thesis, proponents of the non-normative race causation theory endorse a further *priority* thesis, which claims that the causal analysis that constitutes the scientific task should proceed prior to any substantive normative analysis of the discrimination phenomenon. My argument below focuses on separability, since undercutting separability suffices to undercut priority. My thanks to Dmitri Gallow for this formulation of the idea.

To the contrary, it speaks, I think, to the consensus among philosophers that discrimination on the basis of race really is not so mysterious and that it is indeed some kind of adverse treatment caused by race. Three of the four major philosophical monographs on discrimination written in the past 15 years or so pass lightly over the issue, adopting the causal reading when it comes up but not elaborating upon the idea any further. Benjamin Eidelson's definition of discrimination, for instance, features an "explanatory condition," which states that the differential treatment "must have some explanatory connection to [their] possession of a given trait to qualify as discrimination on the basis of that trait" (19). He does not go on to provide a reductive account of what it is, exactly, for the trait to "explain" the discriminatory action, but his discussion implies a causal theory. Kasper Lippert-Rasmussen and Sophia Moreau similarly fill in the "because of" clause in their respective analyses as referring to a causal explanation or causal chain which links the basis of the discriminatory action to the action itself and leave the matter at that.[2] In both cases, what it is for, say, race, to cause the discriminatory action is taken to be fairly self-explanatory.[3]

The social sciences, by contrast, are much more forthcoming about what constitutes discriminatory treatment. Racial discrimination is adverse treatment caused by race; causation, in turn, is a difference-making relation, typically articulated in counterfactual terms. "Measuring Racial Discrimination" gives a clear statement of this standard social-scientific approach: "[T]o measure discrimination researchers must answer the counterfactual question: What would have happened to a nonwhite individual if he or she had been white? Answering this question is fundamental to being able to conclude that there is a causal relationship between race and discrimination, which, in turn, is necessary to conclude that race-based discriminatory behaviors or processes contributed to an observed differential outcome" (77). From here, the National Research Council proceeds to frame the all-crucial counterfactual question in the by now very familiar terms of causal inference: "The causal effect of racial discrimination is the difference between two outcomes: the outcome if the individual were black and the outcome if the individual were white" (79). The causal effect of racial

---

[2] Causation figures centrally in: Benjamin Eidelson's account as the "explanatory condition" of discrimination, in Kasper Lippert-Rasmussen's discussion of the "because of" condition in his section on "Because," and in Sophia Moreau's discussion of what unifies direct and indirect discrimination. See Eidelson 2015, pp. 17–24; Lippert-Rasmussen 2013, pp. 36–40; Moreau 2020b, pp. 19–21. Deborah Hellman's book is the only one that does not discuss causation, since it focuses rather exclusively on the wrongmaking feature of discrimination and so does not consider the conceptual question; see Hellman 2008. Hellman, however, has picked up the question in recent work. Calling it a "definitional" issue, she writes, "Despite the centrality of the concept of disparate treatment [discrimination] over many years, we are lately coming to recognize that we don't quite know what it is." The problem, in her eyes, sets forth "a research agenda for our times" (Hellman 2023, pp. 206–207).

[3] There is also the fact that discrimination has been primarily conceived as a topic for moral and political philosophy, which explains the relative neglect of discrimination's conceptual aspects in comparison with its normative ones.

discrimination is simply the causal effect of race. If race causes some differential outcome, then the differential outcome is due to racial discrimination.[4]

"Measuring Racial Discrimination" defines the effect of racial discrimination as the effect of race *neat*, as it were. Other works in the social sciences define racial discrimination as a more specific kind of causal effect of race. For example, Judea Pearl, widely regarded as the progenitor of the Structural Causal Model (SCM) framework for causal inference, takes the legal prohibition on racial discrimination to correspond to a prohibition on race's having a "direct effect" on some outcome. In his foundational paper "Direct and Indirect Effects," Pearl writes:

> Another class of examples [of direct effects] involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they have on hiring by way of applicant qualification. This is made explicit in the following court ruling:
>
> > 'The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same.' (Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).)
>
> Taking this criterion as a guideline, the direct effect of X on Y (in our case X=gender Y=hiring) can roughly be defined as the response of Y to change in X (say from X= x* to X= x) while keeping all other accessible variables at their initial value, namely, the value they would have attained under X= x* (emph. in original, 2001, 412).

In the Potential Outcomes framework of causal inference, racial discrimination is most often defined as the causal effect of *perceiving* race—i.e., in the case of employment discrimination on the basis of race, the causal effect of an employer's perceiving the job candidate's race. This approach clarifies the causal account of racial discrimination not by specifying the particular type of causal relation that holds between race and some outcome (as, for example, Pearl does above in defining it as specifically the "direct" effect of race) but by sharpening the perhaps too-ambiguous causal variable "race" whose causal effects are under investigation. On this view, the law identifies a particular kind of causal effect of race as racially discriminatory: the effect of *race perception*. Race perception is therefore the cause or, in the language of Potential Outcomes, the "treatment" of interest in empirical investigations seeking to detect whether racial discrimination has occurred.[5]

---

[4] TKTK Cite other prominent social scientific works

[5] This emphasis on race perception as the variable of interest is the joint product of both a popular legal interpretation of "disparate treatment" discrimination and methodological concerns native to the Potential Outcomes framework. James

Still other social scientists endorse the general equation of race discrimination with race causation but are more vague about the particular race-based causal relation that constitutes discrimination on the basis of race. Writing explicitly on the use of statistical (as opposed to explicitly causal) methods to test for the legal concept of disparate treatment discrimination, economist and legal scholar Ian Ayres writes that the aim of such analyses is "to ascertain whether an applicant's race was the cause of being denied employment" (4). To do so, "[d]isparate treatment tests, at least in theory, should strive to control for any and all variables that plausibly had a causal impact on a defendant's decisionmaking" (5). Because Ayres' discussion centers on regression techniques and so does not itself take place within an explicit framework for causal inference, it is difficult to tell the exact kind of causal effect of race he takes to be the hallmark of racial discrimination. But his emphasis on striking out the causal influence of "any and all variables" but race—including variables which are themselves causally influenced by race—is markedly similar to Pearl's dictate that "all other accessible variables" be held fixed. The focus on "whether the race of the plaintiffs had an independent influence on the defendant's decisionmaking" (8) is akin to Pearl's formulation of nondiscrimination as requiring that "race do[es] not *directly* influence hiring decisions" (emph. in original).

These philosophical, social-scientific, and legal characterizations of discrimination on the basis of race as a kind of differential treatment caused by race are exemplary of what I am calling the "non-normative race causation" theory of racial discrimination. Proponents of this approach analyze what race causes in order to analyze whether some differential treatment or outcome constitutes discrimination on the basis of race. Put differently, (mere) differential treatment becomes racially discriminatory treatment when it is caused by race.[6]

It is important to emphasize the contrast between this view of discrimination as caused by race with a more general causal thesis about racial discrimination: that racial discrimination is *some* causal process which has something to do with race, although is not necessarily *caused by* race. Consider, for instance, an alternative view on which some practice is discriminatory on the basis of race if it leads to an increase in racial inequality. This account can be elaborated in terms of causation

---

Greiner (2008) writes that "what matters is not actual race or actual gender (which is usually conceptualized as being set before birth), even if there were a general consensus on what those terms mean. Rather, what matters is the state of the world as of some moment of the social or governmental actor's perception" (577). Gaebler and Goel (2015) write, "Disparate treatment concerns the causal effect of race on outcomes—as we formalize here by the $sate_M$—with behavior often driven by animus or explicit racial categorization."

[6] As Eidelson (2015) puts it, "What distinguishes discrimination from mere differential treatment is its special explanatory connection to some differential ascription of a property to the discriminates" (13).

and in terms of race. Treatment is racially discriminatory if and only if it causes a rise in inequality among those of different racial statuses. But this account does not amount to what I am calling a "non-normative race causation" theory of racial discrimination. For it does not define racial discrimination as race causation, that is, treatment *caused by race*. According to this alternative view, what makes a practice racially discriminatory is not whether or where race figures in its etiology, as race causation accounts claim, but the fact that the practice has a certain racial upshot or effect—an increase in racial inequality—whatever its causes. By contrast, all of the approaches to analyzing racial discrimination mentioned take racially discriminatory treatment to be defined by what brings it about: race.

This race causation strand of theorizing about racial discrimination is *non-normative* in that it presumes that the causal connection between race and the discriminatory treatment can be understood without reference to the moral character of any racial differences. This non-normativity is crucial to the reductive ambitions of the larger projects within which these analyses are situated.[7] If the causal analysis of race fails to be non-moralized, it cannot make any progress towards an analysis of racial discrimination at all, at least as these theorists conceive it. A new approach will be required.

§3 Against non-normative race causation

There are two main ways that authors have gone about the race causation approach to racial discrimination. The first follows what I will call a "wholly negative" approach to defining the target discriminatory effect, which declines to put forth any positive sense of what underlies the effect of interest. Pearl's formulation introduced above is an illustrative example of this approach. These accounts succeed in starting from a normatively-neutral place but cannot go on from there to coherently define the causal effect that is supposedly discriminatory. The alternative is to start from some substantive, and so not wholly negative, account of the discriminatory race effect. This approach succeeds in defining the target effect but only because the account from which it proceeds

---

[7] My focus on non-normative causal analyses of race therefore sets to the side alternative theories on which what race causes is partly defined by moral and political considerations to do with, among other things, racial discrimination. These approaches are entirely absent, as I see it, from the multidisciplinary literature on discrimination. The closest line of work, which explicitly combines causal and normative analysis is "Path-Specific Effects," which defines racial discrimination as some normatively-determined subset of race's causal effects. Still, even here, the causal analysis and ethical analysis are conceived as independent of one another, the latter applying on top of the former. Elsewhere, I have argued that that a normative-causal analysis of race does not necessarily make for a circular theory, but it does block the kind of reductive analysis that these authors are after (Hu 2025).

is not normatively-neutral and so cannot vindicate the separability of the causal and the normative parts of a theory of discrimination. Those who look to analyze racial discrimination by way of racial discrimination are thus faced with a dilemma.

§3.1 The wholly negative approach

It is common to describe causal effects in *negative* terms: as the dependency that remains once the effects of other relevant factors are accounted for. Interventionism can itself be thought of as giving a kind of negative account of causation. The causal effect of X on Y is the difference that changing X makes to Y in conditions where those effects on Y which do not trace back to X are all eliminated. Thus, a process of elimination—not this effect, not that effect, and so on, but what is left over having eliminated that set of effects—captures both how causal effects are theoretically defined within the interventionist framework and how they are statistically identified within the causal inference framework by sifting through masses of purely correlational data and controlling for confounders.

Whereas a standard negative account defines some target effect by picking out a particular set of effects which must be held fixed and struck out in order to leave the effect of interest, what I will call a *wholly negative* account defines the target effect as the effect that remains having held fixed and eliminated "everything else"—where what counts as "everything else" is left unspecified. Pearl's account is illustrative of this approach. Pearl, recall, defines racial discrimination as differential treatment *directly* caused by race. He in turn fills in the notion of a "direct effect" of race on, in the case of employment discrimination, hiring, by describing it as the change in hiring in response to a change in race, while "keeping all other accessible variables" as they are.[8] This definition, Pearl tells us, mirrors the legal notion of discrimination, which is defined in precisely these counterfactual terms: "whether the employer would have taken the same action had the employee been of a different race… and everything else had been the same."

I want to grant for the sake of argument Pearl's interpretation of antidiscrimination law, in fact we can set aside the issue of discrimination entirely for now, in order to focus on this definition of the direct effect of race on hiring decisions. What I want to call our attention to is the wholly negative approach that Pearl takes to defining this effect, rather than the specific kind of effect (the

---

[8] The modifier "accessible" here is clearly not meant to be a substantive restriction on which variables should be held fixed to identify the direct effect. Rather I take it to simply refer to the fact that any exercise of causal inference can only work with the variables at hand. I discuss this further below.

direct effect) that he defines. Does this wholly negative account work to pick out a particular effect of race that might then be the effect of racial discrimination?

First, it is important to see that Pearl's definition of the direct effect really is a wholly negative account rather than a standard negative account defined around a particular set of effects represented within a given model. The claim here is that the direct effect of race on the hiring decision is the effect that remains holding fixed everything else that causally influences hiring—not everything else that happens to be accounted for in some given model. That formulation, after all, would problematically give rise to a model-relative notion of racial discrimination, which would clearly be inadequate for the purposes of adjudicating disputes over whether discrimination in fact occurred in some particular circumstance. To be of use in the context of establishing instances of racial discrimination, causal claims about race cannot depend on a subjective choice of which variables to fix. The idea behind the direct effect formulation seems to be that the direct effect of race is the effect that issues from race *itself*. Thus, all effects that are mediated through other factors must be blocked, in order to leave just the direct effect of race.

As I've said, defining the target effect of X on Y negatively via a kind of process of elimination as the change in Y, which is not due to this effect, not that effect, not the other, and so on is commonplace in philosophical theories of causation and scientific causal inference. There is nothing inherently wrong with this way of proceeding; but the process of elimination must minimally have the resources to determine what to eliminate (i.e., in Pearl's case, the indirect effects) and what not to eliminate (i.e., the very direct effect of interest). To claim simply that "everything else" must be struck out en route to identifying the target effect while declining to articulate any positive account of the target around which "everything *else*" is defined, is to refuse to draw this line. Pearl's wholly negative approach therefore cannot maintain the essential distinction on which it relies, between: (i) the indirect effects, which must be struck out to properly distill the target direct effect of interest, and (ii) precisely that target direct effect. Put another way, the wholly negative approach to defining a causal effect is a purely formal approach. And just as a purely formal interventionist causal analysis cannot get off the ground without a substantive account that fills in what exactly must be manipulated and what exactly must be held fixed or what exactly makes something an effect of interest as opposed to a confounder of that effect so the wholly negative account of the direct effect cannot get off the ground while leaving entirely unspecified what *makes* any effect direct or indirect.

And as impossible as it is for the process of elimination to start up, it is equally impossible for it to wind down. For the process would either lead down a regress of striking out indirect effects until eventually no direct effect remains to be identified at all, or it may be stopped short of completely unraveling but on entirely arbitrary grounds. Consider, for example, how this plays out concretely in the case of the direct effect of race or sex on hiring decisions. Let's grant for the sake of argument Pearl's claim that the effect on hiring that goes by way of race or sex's effects on applicants' "qualifications" is indirect. What are the other indirect effects that race or sex have on hiring? An applicant's race or sex certainly causally influences far more than their qualifications, so this is presumably not the only indirect effect within the "everything else" that must be struck out to identify the direct effect. Race and sex also influence, for instance, what position an applicant applies to, what they wear to the job interview, how they answer questions in the interview, how they would fit in at the workplace, among innumerable other factors. Should these effects not also be struck out of the direct effect of interest?

The general issue here is that the distinction between direct and indirect effects is not just elusive within the modeling framework, it is ultimately illusory, since all causal effects, however "direct"-seeming, can always be characterized as being mediated via some intermediate factor. That is, one can always interpolate some variable Z in between X and Y, and thereby turn any "direct" effect into an "indirect" one.[9] As a result, the process of elimination which looks to cross out all indirect effects has no backstop. So, various mediators between race or sex and the hiring decision are teased out and struck out as indirect effects en route to identifying the direct effect—until the process eventually collapses in on itself, with the result that there is no direct effect of race or sex on hiring after all.[10]

Taken to its logical conclusion, then, the wholly negative account leads in this case to a rather noxious moral upshot: when eventually no dependence remains, there is no direct effect and in turn, no discrimination. Meanwhile, one can only avoid this outcome by carving out exceptions to the process of elimination; but these exceptions cannot be grounded without a positive account of what the direct effect of race or sex on hiring is. So, any attempt to exempt those hiring decisions

---

[9] I've argued elsewhere that a positive objective notion of a direct effect does lurk behind these negative model-relative notions, though they are almost always left implicit and untheorized. (Hu, ms.)

[10] To see why retreating to model-based causal analyses won't work another way, consider that the direct effect as negatively defined in any given causal model is only as good an approximation of the target direct effect to the extent that the model accounts for all the (real) indirect effects. But now what are these? There must be some objective non-model-relative distinction between indirect and direct effects, in order to be able to distinguish better and worse causal models and so better and worse direct effects.

motivated by e.g., sheer racial antipathy or plain sexism—paradigmatic cases of race and sex discrimination—as *not* indirect effects to be struck out of the direct effect cannot be grounded in the wholly negative account. For to refuse to elaborate any positive account of what underlies the direct effect is to rob oneself of the ability to designate any effects as a part of it.[11] Any effort to preserve some specific effect of race as a *part* of the direct effect of interest or strike out the effect of race as *distinct* from it can only be grounded in a substantive account of what this direct effect *is*. This was the central lesson of chapter TK: the distinction between the target effect and a confounder of that effect is a distinction that must be grounded in some minimal account of what positively underlies the effect of interest. That account is the substantive background causal structure upon which interventionist analysis applies—without which causal analysis cannot even get off the ground.

§3.2 A substantive approach

A causal analysis of race which is not wholly negatively defined is one that starts from some substantive theory of what race is and how it fits within some broader causal structure of interest. What kind of causal information is this exactly? Here is it helpful to think of a concrete example. Consider the effect of race (R) on likelihood of being hired at some company (Y). To be able to identify this effect, one must know whether to hold fixed, or "control," some factor Z which is causally relevant to Y, or whether to let Z covary alongside R. In order to know this, one must know whether that factor Z is a part of the causal pathway(s) by which R causes Y or distinct from it. And to know this in turn requires that one know *something* about the way race would cause hiring. In particular, one must know enough about how race might cause hiring to at least be able to distinguish between those effects which are and are not a part of how race causally influences (or would causally influence) hiring—that is, in order to know whether Z's effects count as part of R's effects on Y or as confounders of it.

We can elaborate the case further to draw out more precisely what this prerequisite background causal knowledge is. Suppose that at the company in question, new recruits are hired through a system of employee referrals, whereby open positions are primarily filled by candidates

---

[11] What's more, one might point to the fact that racial antipathy and outright sexism route through, say, "hostility" or "perceived trust" to affect an employer's hiring decisions to argue that they indeed are indirect effects. After all, one might say, an applicant's race or sex causally influences the employer's perception or judgment that they are [unfit for the job/untrustworthy/suspicious/etc.], which in turn causally influences their hiring decision—so they are not direct and thus not discriminatory effects. Now, whether this is the correct or incorrect verdict on employment discrimination is beside the point; rather it is that, having started this process of elimination, the wholly negative account of the direct effect is hard-pressed to avoid this conclusion.

who are referred to apply for the job by existing employees at the company. And suppose, as is in fact the case in contemporary American society, that the social networks from which referrals are drawn are marked by racial segregation. Are the resulting racial differences in hiring caused by relying on this practice, differences caused by who happens to get referred to apply for a job—and not, therefore, differences caused by race? Or are hiring practices which rely on social networks simply one *way* that race influences hiring outcomes—that is, a part of *how* race causes hiring differences? If differential access to the kinds of social connections which facilitate access to employment are a part of how race causes hiring differences—if social network access *mediates* an effect of race on hiring—then it should vary alongside the change in race, so that its effect may be captured as a part of race's effect o[n hiring. Whereas, holding fixed the effect of such network access would otherwise block part of the effect of race on hiring. On the other hand, if such access is distinct from how race influences hiring—if network access is *off-path* of race's causal pathways to hiring—then its effect positively should be blocked so not to confound the effect of race *in particular* on hiring.

In the first part of this book, I argued that the requisite task of outlining this background causal structure of how race acts as a cause of hiring necessarily proceeds from a normatively thick understanding of the social world. Different causal models put forth different theories of how race acts as a cause and how racial difference is produced, theories which are not always directly comparable on purely empirical grounds. Most obviously, different models might conceptualize the race variable differently and accordingly, conceptualize race's causal effects differently. But even where models employ the same conception of race, their construction is underdetermined by the empirical data. As I argued in chapter TK, models are often either empirically indistinguishable, because they equally fit the observed data, or incommensurable, because the immense complexity of the social world means that every causal theory of race will invariably be partial and so often not directly comparable with others. How a model is constructed is not then dictated by empirical considerations alone. Rather, a model of how race acts as a cause reflects a normative social theory of how racial differences are produced and reproduced. Implicit (or explicit) in that theory are moral and political judgments on such matters as what sources of racial inequality are most significant and how to conceptualize responsibility for racial injustice. Causal models of race thereby become laden with these moral and political values.

Thus, whether in the hiring case described above, network-based referrals should be held fixed or "controlled" when identifying the effect of race depends on this broader normative

understanding of how race acts as a cause of hiring. It depends, for instance, on how the racial composition of employee referrals relates to race. Is race a cause of who gets referred? Or are referrals caused by some factor which is merely correlated with race? Insofar as an analysis of the causal effect of race on hiring entails making a choice to either hold fixed or not hold fixed network-based referrals, it cannot sidestep these questions. But these requisite questions cannot be settled by empirical data alone. And thus insofar as the answers to these questions will rest on both causal-empirical and normative considerations about race and hiring, the causal analysis of the effect of race on hiring will itself rest on both causal and substantive normative considerations about race and hiring—against the separability thesis.

Recall at this point the role that these causal analyses of race are meant to play within the broader philosophical, legal, and social-scientific analyses of racial discrimination. Their purpose is to set to the side any debate about the normative dimensions of discrimination and characterize the phenomenon in purely descriptive or empirical terms. Clearly, the inevitable entry of normative judgments about race and racial difference into the causal analysis undercuts this ambition. But even more worrisome is what this means for the larger project of theorizing racial discrimination via race causation. It is one thing for a theory of race qua cause to embed substantive normative judgments about race, racial differences, and racial injustice. It is another thing for it to embed substantive normative judgments about racial *discrimination*, that is, the very phenomenon at issue. But is it really plausible to think that the former set of normative judgments do not at all touch on the matter of discrimination? There are good reasons to think that normative judgments about race and racial difference do involve judgments about discrimination. Insofar as racial discrimination is not a sui generis category of racial wronging and harm but rather derives its normative significance from broader issues concerning race and racial difference, judgments about the latter cannot be cordoned off from judgments about discrimination.[12] Moreover, discrimination as a legal concept is closely tied to antidiscrimination legal tools, thought by many as aiming to achieve broader societal goals of redressing racial inequality. From this vantage, normative judgments about the kinds of racial inequality which ought to be redressed figure in judgments about the scope of antidiscrimination law and so in turn, what constitutes racial discrimination.

To fail not just to be normatively-neutral but to be neutral with respect to moral and political views of racial discrimination is to really flout the vision that underlies the separability thesis: that

---

[12] TKTK cite

while there is a technical task here to be carried out by experts specially trained in the science of hunting causes and effects, that job should be marked off from and prevented from encroaching on what ought to be a public ethical debate about what constitutes wrongful or unlawful discrimination. The violation of this kind of separability has serious consequences. First, I take it that much of the plausibility of the non-normative race causation approach to racial discrimination derives from the claimed preservation of some special space for normative thinking about discrimination. Without it, I suspect that the approach would be received with much more skepticism. Second, and to which I will now turn, the impossibility of separability raises important normative questions about both the analyses of discrimination produced by the race causation approach and how those analyses are produced.

§4. Normative consequences of a *normative* race causation approach to discrimination

I have argued that the proponent of the race causation approach to analyzing racial discrimination is faced with a dilemma.[13] If she tries to eschew all substantive theorizing about how race acts as a cause (that might introduce normative considerations into her work), she finds that she cannot even get her causal analysis off the ground. If she allows herself to start from a positive account of race qua cause, she finds that her causal analysis of race rests on a background social theory and series of methodological choices which are grounded in not just empirical but normative value-laden judgments about race. In either case, the prospect of a non-normative race causation theory of racial discrimination lies out of reach.

Of course, the impossibility of fully achieving the clean division of labor envisioned by the separability thesis does not impugn the ideal itself. It remains intuitive to think that nailing down what happened as an empirical matter should be prerequisite to judging whether what happened was wrong or unlawful. An approach to discrimination that explicitly separate causal analysis from normative analysis based would thus seem able to draw on specialists' technical expertise while reserving moral and political reasoning about race and discrimination for a wider non-expertise-delimited populace. Breaking down the complex inquiry into smaller and more well-defined parts—

---

[13] Having argued that the predominant approach to analyzing racial discrimination by way of race causation fails to be normatively neutral in the way its proponents hope, I will from here on out be mostly dropping the "non-normative" modifier to this race causation approach.

causal facts here, normative judgments there—would also serve to make the decision-making process more transparent and procedurally fair.

But the extent to which these virtues are realized depends, of course, on the extent to which separability is realized. If in the course of analyzing race's causal effects, experts necessarily cross back and forth between causal thinking and value-laden normative thinking about race, as I have argued, then we have reason to doubt that inquiry into discrimination organized around this (ostensible) division of labor really is more rationally organized, transparent, and procedurally fair. We should be skeptical that the sharp boundaries imagined by the separability thesis do in fact serve to protect and preserve the space of ethical reasoning reserved for non-experts, as proponents of the race causation approach thought it would. What is more, it is not merely that these values which we would have secured are, sadly, unable to be. So long as analyses of discrimination continue to operate on the *model* of separability, they serve to sustain an illusion of these virtues when they are not in fact present. Meanwhile, causal analyses and their practitioners, which I will call "causal experts," retain a privileged role in analyzing discrimination, a role that molds not just our public understanding of discrimination but, just as importantly, our ability to collectively shape that understanding. Given the inability to wholly separate causal and moral and political theorizing about race, this is a level of influence that far outstrips its justification.

It will help to divide these normative consequences of the impossibility of separability into two categories. There are first those consequences that have to do with the resulting accounts of discrimination that are produced by normative race causation approaches. These are ones in which the special role awarded to causal experts grants their moral and political views a privileged place, too. The expert thus has outsized influence in determining paradigmatically normative issues in discrimination than the non-expert.

The impossibility of separability also has broader implications beyond its effects on the resulting account of discrimination. First, the non-normative race causation approach to analyzing discrimination dictates that we think first about what race causes (without drawing on any substantive normative resources for doing so) and then think about whether the effect is wrongfully discriminatory. This is a methodology that gives a kind of priority to causal theorizing over normative theorizing, to distorting effects on how we reason about discrimination; the impossibility of separability makes these effects particularly pernicious. Second, to the extent that the promise of value-neutrality expressed in the separability thesis lends credence not just to the conclusions of race causation approaches to analyzing discrimination but to those institutions that produce those

conclusions, finding that no such neutrality is possible threatens to undermine trust in those institutions themselves. On the other hand, continuing to falsely believe that the standard race causation analysis is normatively neutral is also troubling; for it serves to shore up the legitimacy of a procedure with the abovementioned antidemocratic implications.

I will call the former set of normative consequences, which have to do with the effects of non-separability on the resultant analyses of discrimination produced by the race causation approach *first-order* consequences. The latter set of consequences are *second-order*: they have to do with the effects that the impossibility of separability has not on the discrimination concept itself but on the broader conditions within which we as a polity come to discuss, deliberate, and determine what discrimination is. This way of dividing things will no doubt fail to neatly sort all of the normative consequences that I will discuss below. Still, it should suffice to highlight the fact that what is at issue here is not just the effects that the race causation approach has on our final judgments about what is or is not discriminatory, but on how we come to make those judgments and the institutional contexts within which we do so.

§4.1 First-order consequences

According to the division of labor set up by the race causation approach, inquiry into racial discrimination should start first with (non-normative) causal theorizing about race, in order to determine the scope of effects that are plausibly racially discriminatory, and then proceed onto moral theorizing about fairness, justice, or whatever other values animate antidiscrimination norms, to determine what kinds of causal effects of race are in fact wrongfully discriminatory and what kinds are permissible or even obligatory. I have argued in this and previous chapters that a causal analysis of race rests on a normatively thick understanding of how race acts as a cause and how racial differences are produced. Because the causal structure within which race operates is highly complex and so is massively underdetermined by empirical facts, a causal analysis of race inevitably draws on a host of background theorizing in order to fix a particular account of how race acts as a cause. I argued that an account of how race is causally operative reflects in particular background moral and political judgments about, for example, the justice of various institutions and the scope of responsibility for racial inequality. By informing the causal model from which causal analysis of race proceeds, these judgments form in part the basis of conclusions drawn about race's causal effects. These causal findings thus rest on both causal and normative judgments about race.

Those who set forth these analyses therefore have their moral and political judgments incorporated into accounts of racial discrimination as a part of their expert contributions to causal theorizing. This is problematic for two reasons. First, insofar as their special input is justified on the specific epistemic grounds that they have special technical knowledge and expertise on matters of causal inference, it should also be circumscribed by those grounds. Analyses which take experts past their area of expertise are less likely to be accurate and reliable than those squarely within their expertise; and so they are less justified in occupying their privileged position. This suggests that the role that causal analyses of race play within discrimination theory more broadly should be appropriately calibrated to the fact that they draw substantially on moral and political judgments about race, which lie outside of causal analysts' expertise. But to the extent that such analyses continue to be perceived as standard exercises of causal inference which sit squarely within analysts' area of expertise, they wield a greater influence within discrimination theory and practice than is in fact warranted.

Second, the privileged role that causal analyses of race play in the race causation approach to theorizing discrimination is not only less epistemically justified than is typically assumed; it effectively grants a privileged role to causal experts' moral and political views, too. The model on which the causal task is carried out before the normative task has the causal analyst setting the stage, as it were, for the subsequent moral analysis by determining what kinds of effects are even eligible for further normative evaluation. But if separability fails and substantive normative judgments play a role at the outset throughout the causal analysis, then the causal findings that are to be evaluated only later have in fact already been shaped by the substantive normative views of causal analysts.

The concern here is that the breakdown of separability—but a continued commitment to the race causation model of analyzing discrimination—yields anti-democratic implications: that the expert in causal analysis will end up systematically exercising greater influence than the non-expert in determining paradigmatically normative issues in discrimination. First, because the causal analysis step affords experts a privileged channel through which their substantive moral and political judgments can shape the definition of discrimination, the causal expert will straightforwardly have greater say, as a quantitative matter, in determining the normative content of discrimination than is democratically justified. Second, insofar as the earlier causal inquiry delimits the scope of the latter explicitly normative inquiry, there is an important way in which the indirect normative theorizing that is an implicit part of the former task plays an even more significant role in determining the normative content of discrimination than the explicit normative reasoning that takes place

downstream of it. At the extreme, the implicit normative judgments embedded in experts' causal analyses can so seriously constrain later normative inquiry as to effectively preempt it. When answers to such foundational normative questions as what the scope of antidiscrimination laws and norms should be or how racial discrimination relates to racial inequality and injustice more generally are already tacitly presupposed in the models from which causal analysis proceeds, much of the substantive normative work has already been done, and little remains for the later dedicated stages of normative analysis to determine.

§4.2 Second-order consequences

The impossibility of separability does not just mean that the race causation approach to theorizing discrimination fails to operate in the way its proponents intended—to various deleterious effects on how the process issues verdicts about discrimination. Understood more expansively, the race causation approach prescribes a general method for reasoning about discrimination whereby causal inquiry about race's effects precedes normative theorizing about what kinds of causal effects are (wrongfully) discriminatory. In the last section, I noted some anti-democratic implications of this methodology, focusing on the difference between how causal experts and non-experts are able to influence the normative content of the discrimination concept. But the methodology does more than just designate particular roles for particular parties in inquiry about discrimination. It structures how we reason about discrimination more generally. Foregrounding matters of causation does not just award causal experts a first-mover advantage in setting the terms of debate about discrimination; it installs the overall rules of engagement that govern discourse about discrimination. In this regime, debate over what constitutes discrimination on the basis of race can easily devolve into a dispute about causes and effects, since the causation element sets a minimal threshold for discrimination. A finding that race did not in fact cause the outcome at issue suffices to trump any claim of racial discrimination; and to respond to the casual conclusion with a moral objection would seem to commit a category error.

A concrete example will be illustrative. Consider a finding that racially disparate stop and frisk rates are not evidence of racial discrimination because they are not in fact caused by race, as shown by some complex exercise of causal inference.[14] The point that systematically uneven exposure to stop and frisk by police according to race is in and of itself demeaning, or that it gives

---

[14] Cite TKTK

racialized individuals a reasonable sense that they are not equals within the society, is irrelevant to the question of racial discrimination, insofar as it does not respond to the causal claim.[15] If the study finds that race did not cause the observed differences in stop and frisk rates, then the discrimination claim does not even get off the ground, the existence of reasonable moral arguments notwithstanding.

This is not to say that the moral objections could not be formulated so as to be directed at the causal analysis. Since a causal analysis of race inevitably relies on a normatively thick understanding of how race acts as a cause, it is liable to being challenged on the basis of that understanding and, in particular, the moral and political judgments that constitute it. In causal inference studies of policing, it is standard methodological practice to distinguish the effect of (perceiving) race on whether an officer decides to stop an individual from the effect of determining whether they are engaged in suspicious behavior.[16] These analyses effectively treat race perception as independent of perception of suspicion, such that differences in perception of suspicion are non-racial and so non-potentially discriminatory grounds for stopping and frisking a civilian. It is important to see that this assumption rests on empirical and normative grounds: first that the two perceptions—perception of race and perception of suspicion—are indeed sufficiently distinct such that one can and should control for the latter to determine the effect of the former, and second, that because much street policing is discretionary in nature, and because the ability to make judgments about suspicious behavior is central among the discretionary powers of police, a causal model of police behavior should allow judgments of suspicion to be a sound basis for decision-making. These two points are by no means self-evident, and the latter in particular may certainly be challenged on moral and political grounds. Still, disputing causal analyses is an onerous task. Causal inference is a highly complex technical exercise, and it is exceedingly difficult, even for experts, to identify the array of background assumptions that support some set of causal conclusions. The prospect of a non-expert's understanding the details of an analysis well enough to pinpoint where a specious normative assumption is being made is highly unlikely. Then there is the further hurdle of getting such challenges from non-experts to actually be taken up seriously by experts.

More importantly, my point is that the mere fact that moral reasoning about discrimination must, as it were, route through causal analysis is itself a mark of the problematic discourse-shaping effects of the race causation model of theorizing discrimination. Even if the complaint can be

---

[15] See Hosein (2018).
[16] See Fryer (2019) and related controversy.

framed as taking issue with a step taken within the causal inference exercise, this is simply not the only or even most natural way of expressing it. Rather, there is something at best confusing and at worst smug and domineering about the requirement that moral and political argument be translated into terms appropriate to causal inference in order to gain a hearing. Such forced translation is expressive of the unequal standing of causal experts and non-experts in determining the meaning of discrimination. Stringent rules of engagement dictating who can speak on what and how might have been justified were separability possible and the clearly demarcated division of labor between causal inquiry and normative inquiry able to be realized. But without it, the mandate to speak in causal talk to be heard is unjustified and puts those not well-versed in causal analysis in an invidious position. The substance of their complaint is perfectly admissible into deliberations about discrimination. It is not in fact invalidated by the results of the causal analysis. Still, it is permitted to enter the discussion only when communicated in the proper language—a language which is foreign, difficult to penetrate without special training, and by its nature overwhelmingly favors experts over non-experts.

One might object at this point that these anti-democratic discursive effects are not a necessary result of the race causation approach, even if it does fail to be non-normative in the way that its proponents intend. Broad recognition that causal analysis features substantive normative judgment might instead lead conversation in the opposite direction: experts may be led to defend explicitly the various assumptions which guide their causal analyses and do so in ordinary language accessible to everyone. In his influential article "Causal Inference in Civil Rights Litigation," legal scholar and statistician Jim Greiner suggests precisely this when he writes that the "critical choices required to draw causal inferences in civil rights litigation are not mathematical. Instead, they depend on decisions about the law and on an understanding of how the world works. These are matters about which judges, lawyers, and laypeople can speak as intelligently as those trained in quantitative methods" (538).

On this reading, the impossibility of separability might even have a salutary democratic effect on how we collectively reason about discrimination. For the specialized methodology that governs causal inference which had previously been the sole province of experts would now have to be justified more broadly. Expert analyses would thus become more responsive to the public, while non-experts' contact with expert analyses facilitates a fuller understanding of the whole inquiry into discrimination. The result is a more democratic science of causal inference and more democratic engagement with the methods by which we come to determine what discrimination means.

I do not wish to deny the possibility that unmasking causal inquiry as relying on substantive background assumptions might push towards a more rather than less democratic model of reasoning about discrimination. But this of course requires that this fact be revealed: that those assumptions become widely understood by judges, lawyers, and laypeople to be contestable and that judges, lawyers, and laypeople are in fact able to contest them. For reasons mentioned above, I am skeptical that either of these conditions are easy to obtain. To the extent that most people remain either largely unaware that causal expertise is open to these sorts of challenges or are, for whatever reason, unable to challenge it, the rosy prospect of such open channels of communication between experts and non-experts is, in my view, unlikely.

In any case, even if it were widely recognized that substantive normative judgments are part and parcel of any analysis of the causal effects of race, it still seems to me implausible that it would usher in the kind of collaboration between experts and non-experts imagined. Indeed, it seems to me just as likely to foster suspicion—that the experts are not truthfully reporting the assumptions which underlie their methodological moves or that they have buried their moral and political judgments in hard to find places—rather than cooperation. After all, the admission of moral and political values in causal analysis constitutes a serious transformation in the image of the race causation theory of discrimination, and the response imagined seems to me an overly neat and circumscribed response to it. That is to say, if promise of value-neutrality expressed in the separability thesis is what underwrites the race causation approach to inquiry about discrimination, then widespread recognition that separability is false would seem to trigger a much more drastic shift in how we analyze discrimination than simply changing how methodological choices within causal inference are justified.[17]

We might reasonably worry, for instance, that widespread recognition that causal analysis is not in fact normatively-neutral as advertised would in turn generate widespread distrust in not just experts and their analyses but moreover the institutions within which they operate—that is, not just expert witnesses and social scientists but discrimination trials and social science writ large. Indeed,

---

[17] One might wonder why Greiner is more sanguine about the possibility that experts and non-experts might productively work together to arrive at the best causal findings. Greiner's optimism derives in part, I think, from his view that the delicate judgments that must be made in the course of using the Potential Outcomes framework to identify causal effects of race (or sex or whatever other protected category is at issue) are mostly empirical in nature. It is one of the primary aims of this book to debunk this widely held assumption. Causal analysis of race proceeds from a thickly normative understanding of how race acts as a cause, an understanding that cannot be fleshed out in purely empirical terms. If that is right, then we should expect the various choice points in the course of causal analysis to be much more controversial and hotly contested.

there is already some measure of distrust that expert witnesses are not neutral arbiters of fact but in fact paid partisans.[18] Now, whether such erosion of trust in the institutions of law and science would in this case be warranted is a further issue, which I will not address here. But insofar as an institution's legitimacy is in part constituted by the people's perception of it as legitimate, serious loss of trust in the institution, whether warranted or not, implicates a loss of legitimacy.

It is no doubt reassuring to think that the race causation approach to analyzing discrimination can serve to strengthen the democratic character of how we reason about discrimination, even when causal experts occupy a privileged position in influencing the normative content of discrimination. But if I am correct in thinking that the appeal and justification of race causation approaches depends crucially on the truth of the separability thesis, then to discredit separability is to discredit the model of discrimination which relies on it. So long as the race causation model persists, the priority it accords to causal inquiry will continue to structure how we reason and deliberate about discrimination, to antidemocratic effects.

---

[18] See e.g., Paul Meier, "Damned Liars and Expert Witnesses" (1986)