HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Philosophy

have examined a dissertation entitled

Causation in the Social World

presented by Lily Hu

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature Edward Hall

Typed name: Prof. Edward Hall

Signature Lucas Stanczyk

Typed name: Prof. Lucas Stanczyk

Signature Tommie Shelby

Typed name: Prof. Tommie Shelby

Date: May 3rd, 2022

Causation in the Social World

A dissertation presented

by

Lily Hu

to

The Department of Philosophy

In partial fulfillment of the requirement

for the degree of

Doctor in Philosophy

in the subject of

Philosophy

Harvard University

Cambridge, Massachusetts

May 2022

Dissertation Advisor: Professor Edward Hall Lily Hu

Causation in the Social World

Abstract

Causation and causal claims abound in the social world as much as in the natural world. But a dominant theory of causation, prevalent among philosophers of causation and scientists who pursue causal inquiry, an *interventionist* theory of causation, is flawed when applied to cases of social causation. Or so I argue. I pursue in this dissertation various challenges to interventionist analyses of social causation: In Chapter 1, I argue directly against a causal structure of the social world limned by interventionism. I claim in Chapter 2 that causal theorizing about social categories such as race involves ineliminable substantive moral and political considerations, a feature for which interventionism cannot well account. In Chapter 3, I turn to discuss the distinctively normative set of issues with adopting a certain kind of interventionism-based causal analysis of discrimination. My suggestion is that these normative upshots make for a strong practical case against the interventionist causal account of discrimination. If I am right in my arguments throughout these chapters, interventionism is not well-suited to the explanatory and normative aims of our causal theorizing. Thus, if the interventionist idea does in fact undergird many of the ways we pursue causal inquiry about the social world, then such practices should either be revised or perhaps even be abandoned.

Table of Contents

Acknowledgements

Thank you to my committee: Ned, Lucas, and Tommie for their guidance these past years of so many twists and turns. To Ned, especially, who always made me feel that I had something to contribute, even when I (as I did frequently) slipped into all-consuming doubt about whether I could get anything right let alone right and interesting. Your brilliance as a philosopher is only matched by your kindness, patience, and generosity as a teacher and person.

Thank you to my friends and family who have carried me through these many years in graduate school, especially: Laura Adler, Danielle Carr, Gabriella Fee, Jessica Fields, Ben Green, Zoë Hitzig, Kevin Hong, Ben Sobel, Kate Vredenburgh, and Zach Wehrwein.

To Salomé: I feel so extremely lucky to have someone in my life who makes me so immensely proud, as though I'm a teen, to call my "best friend."

To Will: for being the most patient, selfless, insightful, supportive reader, interlocutor, editor, critic, etc. etc. anyone could ask for, for nearly a decade.

Finally, this dissertation would not be possible without Issa Kohler-Hausmann, who is equal parts my intellectual twin, partner, and hero. Over hundreds of hours discussing our shared causal obsession, it has become impossible to discern which ideas were first "yours" or "mine"; they have fully melded into a single undifferentiated mass (of genius) that I have drawn from for ideas and inspiration for this dissertation and for much of my work beyond. I can safely say there is not a single person on this planet who is more wholly consumed and also shaken to their core about causal inference methodology, and I feel so fortunate to have you alongside me through it all, just two denigrated causal artists on this long journey of the life of the mind.

Introduction. The Interventionist Idea About Causation

For a causal interventionist, X causes Y just in case it is possible to intervene to change the value of X to change the value of Y. As a first pass, X and Y here are variables, each representing some feature or component of the causal system under study, its set of values corresponding to states that that feature or component of the system is in. (More on variables and variable values shortly). An intervention is an exogenous manipulation of some causal system that makes a change to the value of its target variable—and leaves unperturbed the state of all other causal factors on scene such that any change that may result in the outcome variable can be traced back to the causal influence of the intervened upon variable. The image of divine intervention gives an intuitive picture of the theory: God swoops into some state of affairs at some designated time *t* and with surgical precision changes how X is, leaving everything else on scene pristine exactly as it was. Then, letting the passage of time take over, we check at a later time *t'*: does Y change? If it does, then X is causally relevant to Y; if not, then not. The more prosaic scientific controlled experiment gives another helpful gloss. Imagine an idealized experimental setup. Does performing an unconfounded manipulation on X result in some change in Y? The manipulation on X is unconfounded if it does not muddy the experimental waters by changing other factors in the setup that might exert their own causal influence on Y in a manner that does not go through the change to X.

The close ties between an interventionist theory of causation and scientific practice doesn't stop there. The interventionist's causal relatum of choice, the variable, is also the scientist's. Broadly conceived, a variable is simply any determinable of the causal system that takes on some determinate, which corresponds to its value. The metaphysician might now want to press further: what *are* variables? Can variables really represent *anything* in the causal system? But just as the scientist pays little heed to the question, the causal interventionist too, by and large, waves it away

and stands firm in her refusal of "metaphysically portentous" constraints.[1] Variables can stand in for really any property of the system, with the only constraint being that its values or settings are mutually exclusive. So long as such a variable can vary its value, it can be a cause or can be affected by one.

Whether a particular manipulation constitutes an *intervention* depends on what the operation does to its target variable X and what it doesn't do to the other causal variables that link up to Y. This, in turn, depends invariably on what exactly these other variables in the system are. Possible interventions and eligible variable sets, therefore, go hand-in-hand: the set of variables put forth to carve up a given causal system constrains the set of operations that meet the technical requirements of an intervention; the set of possible manipulations (with an account of what constitutes "possible" yet to be fully spelled out) constrains the variables and sets of variables that may represent a given situation. This then yields one constraint on variable construction the interventionist cannot shake off: if their role in some causal structure is to be illuminated by interventionism, variables must be eligible targets of interventions.[2]

While it goes without saying that the core concept in an interventionist theory of a causation is the intervention, the place of eligible variables and variable sets in interventionism has been much less remarked upon—despite this "consistency" constraint, as Woodward puts it, between variable sets and local interventions.[3] An interventionist's starting kit of variables determines the kind of causal structure she will be able to trace out with her analysis: the range of alternative states of the system under consideration, the "level" or "scale" of her causal analysis, whether physical, psychological, or social features of the system are highlighted, and so on. Naturally then, the

[1] James Woodward, "Mental Causation and Neural Mechanisms," in *Being Reduced: New Essays on Reduction, Explanation, and Causation*, ed. J. Hohwy and J. Kallestrup (Oxford: Oxford University Press, 2008), 218–62, 231.
[2] Here I am highlighting *relational* rather than non-relational constraints on variable construction. David Danks first made note of this distinction.
[3] James Woodward, "The Problem of Variable Choice," *Synthese* 193, no. 4 (2016): 1047–1072, 1062.

goodness of any resulting account of causal structure will depend in considerable part on these variable building blocks. An analysis of variable choice, of what variables in a causal system can and should be, stands as central to an interventionist theory of causation as an analysis of the technical notion of an intervention.

In philosophy, the interventionist account of causation was powerfully set forth by James Woodward in his 2004 *Making Things Happen*, though the core idea of the theory has been in circulation among theorists of causation for much longer.[4] Computer scientist Judea Pearl began developing in the 1980s a set of formal tools for causal inference based in directed acyclic graphs and the idea of intervening to set variables to take certain values, culminating in the publication of *Causality* in 2000.[5] Around the same time, philosophers Peter Spirtes, Clark Glymour, and Richard Scheines published *Causation, Prediction, and Search*, which too centered a graphical approach alongside the notion of an intervention in an analysis of causation.[6] And precursor to both projects was the effort by economists in the first half of the 20th century to set a framework for proper causal interpretation of models and observational data based on hypothetical manipulations.[7] These more technical explorations of the interventionist idea found more solid philosophical footing when they linked up with David Lewis's counterfactual theory of causation and formulated interventionism as giving a kind of counterfactual analysis of causation.[8]

---

[4] James Woodward, *Making Things Happen* (Oxford: Oxford University Press, 2003).
[5] Judea Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press, 2009 [2000]).
[6] Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, Prediction, and Search* (Cambridge: MIT press, 2000 [1993]).
[7] See e.g., economist Trygve Haavelmo's work on causal interpretations of economic models grounded in hypothetical manipulation published in two seminal papers in 1943 and 1944 and the work of the Cowles Commission, a research program founded in 1932 that greatly influenced modern econometrics, that formalized causal interpretations of empirical relationships. Trygve Haavelmo, "The statistical implications of a system of simultaneous equations," *Econometrica* 11, no. 1 (1943): 1–12; "The probability approach in econometrics," *Econometrica* 12, Supplement, iii–vi and 1–115.
[8] David Lewis set forth a counterfactual program for analyzing causation in the 1970s in "Causation," *The Journal of Philosophy* 70, no. 17 (1973): 556–567 with revisions to his theory in "Causation as Influence," *The Journal of Philosophy* 97, no. 4 (2000): 182–197.

Shared among proponents of interventionism is a broadly pragmatic view of causation: that a good account of it will make for a concept that will prove useful for our various theoretical and practical projects, and in so doing, show why we use causal concepts so widely in both our everyday lives and in our more specialized endeavors—hence many interventionists' motivations to develop an account of causation that fits well with our best scientific practices in causal inquiry.[9] An adequate theory of causation will therefore do more than simply predict accurately. Our sciences are after all interested in explanatory structure, something which goes above and beyond mere predictive success. Furthermore, in harmonizing with scientific inquiry, a good causal theory will seek a balance between rationalizing existing practices and guiding towards better ones. Interventionists are interventionists because they take it that an analysis centering difference-making under surgical interventions makes for an account of causation that does best when measured up against these ends. It presents an analysis that best meets our causal cognitive needs, or best accords with and rationalizes practices of scientific causal inquiry, or is most fruitful for our causal-explanatory projects. These virtues make it so their theory wins out.

I want to now frame this point about interventionism's ability to make good on these aims of a theory of causation in a way that brings out the problem of variable choice. On one hand, there are those variables and sets of variables that are "good" as regards those ends: e.g., variables that have great explanatory potential, or that have a prominent place in our scientific inquiry, or that our folk intuitions readily and firmly identify as causal, and so on. On the other hand, there are those variables and sets of variables that interventionists can accommodate within their theory, where,

[9] Contrast this view with one that looks primarily to track our folk intuitions about causation. Our practices of scientific causal inquiry do not starkly contrast with our intuitions about causation or our everyday use of causal language. Hence although different causal theorists will emphasize different aspects of these success criteria, competing theories of causation are not just talking past each other. Woodward has clarified the aims of his interventionist theory of causation in much of his writing since *Making Things Happen*. Two particularly notable articles on the matter are "A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment)," *Philosophy of Science* 81, no. 5 (2014): 691–713 and "Methodology, Ontology, and Interventionism," *Synthese* 192 (2015): 3577–3599.

recall, even those interventionists who cast a wide net on variables must commit to the constraint that her variables be candidates for surgical interventions. To hear Woodward on the matter, variables ought to be

> well-defined targets for (single) interventions in the sense that they describe quantities or properties for which there is a clear answer to the question of what would happen if they were to be manipulated or intervened on. One obvious rationale (again within an interventionist framework) is that to the extent this requirement is not met, a representation employing these variables will not provide answers to questions about what will happen under interventions on such variables and hence will fail to provide information about causal relationships.[10]

Interventionists who are functionalists about causation are interventionists because they take it that variables and sets of variables which satisfy both criteria for variable construction exist. That is, there exist variables that are consistent with the technical notion of an intervention *and* that we in our everyday lives and in our scientific inquiry care about. Upon resolving this problem of variable selection, the interventionist can then provide an account of causal structure that does well with respect to the ends of her causal inquiry. The task, and indeed it remains a task for the interventionist to figure how best to carry out, is to construct the set of variables that can perform this double duty.

But what if such variables do not exist? What if "good" variables in some domain vis-à-vis our causal practices are not "well-defined targets for (single) interventions"? Where would this leave the interventionist? One, I take it, flat-footed response would be to deny these variables as being causal or explanatory after all, saying: "All the worse for our scientific practices!" Interventionist appeals for good hygiene in our causal talk are certainly fair enough, though there is a fine line here between calling for clarity in our causal claims and insisting that every claim must fit the strictures of interventionism. Requiring a retraction of all those factors that cannot easily fit the interventionist paradigm seems to me to amount to little more than doubling-down on the matter of whether we

[10] James Woodward, "The Problem of Variable Choice," 1054.

ought to analyze causation in interventionist terms. It is to simply beg the question in favor of the theory.

If variable sets come intertwined with possible interventions, then the problem of variable selection becomes a problem of the intervention. And failing to successfully revise her core concept, the interventionist cannot provide a perspicacious analysis of causal structure, at least not in this particular area. The interventionist who looks to illuminate the causal structure of the *social world* finds herself in precisely this unfortunate position. Or so I will argue.

The argument for how and why the problem of variable selection poses such a challenge to an interventionist analysis of social causation unfolds over the course of the subsequent chapters of this dissertation. But I will here preview how a variable can fail to be properly subject to an intervention. The claim might seem at first quite mysterious. If all variables must be constructed so as to be able to take on one of its (exclusive) values, then how could a variable fail to be the target of some hypothetical intervention that changes its setting from one to another? What kind of a thing would fail to bend to the surgical precision of God's hands?

To approach the question, we must revisit the technicalities of an intervention. Recall the main idea: an intervention on X with respect to Y changes the value of X and makes no direct change to any of the other variables in the causal system. All other variables are therefore held fixed at their actual values, unless they are changed *by way of* the change made to X. If, under these conditions, the value of Y changes, then X is causally relevant to Y. Interventionists take this to be the idea that underlies the idealized randomized controlled experiment in our scientific inquiry. An experimental setup that randomizes administration of some drug disentangles the causal effect of the drug from the causal effect of other factors that might matter for an individual's health outcome. The medical intervention determines only whether you are treated with the drug or not, and has no

effect on, for instance, your diet and exercise regimen or whether you have a particularly strong immune system—lest observed differences in health outcomes be due to these factors, which are distinct from the drug's causal effects but happen to correlate with the drug treatment assignment. It is because randomization is taken to approximate the ideal intervention that the randomized controlled trial is taken to be the gold standard of causal inquiry.

The causal interventionist, unlike the experimenter, need not worry about what it would take to actually perform an intervention on some variable X. She need not even be concerned about whether X is actually manipulable. What matters is that an intervention that changes a target variable's value is possible in a very wide sense of possibility, perhaps logical possibility, or as Woodward suggests, possible in the sense of not being "ill-defined for conceptual or metaphysical reasons."[11] So, the kind of variable that is likely to fail to be the target of an intervention is one that cannot be changed without perturbing, *in the same fell swoop* of manipulation, other causal variables that link up to Y via causal paths of influence that do not proceed from the intended target of the manipulation. A variable, that is, for which there exists no possible manipulation that does not also directly change the values of other causally significant variables. This condition on eligible variables and variable sets appears deceptively minimal. However, as we will see in the subsequent chapters, the idea of an intervention in fact sets considerable constraints on variable construction, ruling out many which would on their face seem to be wholly unproblematic and more, which we want to study from a causal perspective.

Here is an example of such a case, which will feature throughout this dissertation. A job interviewee receives a poor interview rating and wants to know what causally explains the outcome. He wonders: was it his perceived *sex status* as male that swayed the interviewer's, or what is it, since

[11] James Woodward, *Making Things Happen*, 132.

he wore a skirt and facial makeup to the interview, his *gender non-conforming presentation*, or was it neither of these factors? Perhaps the interviewer harbors a resentment against all interviewees she takes to be sex-coded male? Or is it that the interviewer is put off by individuals who present as gender non-conforming? The interventionist tries her hand at the causal query, but she quickly runs into trouble. A manipulation of the interviewee's sex status (that holds fixed the interviewee's dress and facial makeup as they are) entails a change to whether the interviewee presents as gender conforming. There is no possible unconfounded manipulation of sex status that disentangles it from gender conforming status as requires the interventionist. Even the precision of God's hands cannot thread this needle.

Here is a different case, this one due to Markus Eronen.[12] Does having *pessimistic thoughts* cause *difficulties in concentrating*? An interventionism-inclined psychologist pursues her inquiry by considering a hypothetical intervention that alters just whether an individual has pessimistic thoughts and changes nothing else about her other psychological states. The intervened-upon individual must therefore not be different in, say, her overall feelings of anxiety or sadness, lest a change to these factors influences her ability to concentrate. But what could such an intervention be? I do not of course mean practically-speaking*;* the problem is rather conceptual. For what it is to have *pessimistic* thoughts is for one to think negatively about one's circumstances. It is to face the world fatalistically, with an orientation towards despair and hopelessness. To think pessimistic thoughts is simply to be in a frame of mind characterized by feelings of deep sadness. No intervention can tease these apart either. And yet the interventionist should have no problem with the case, for it is quite clear what the intervention is supposed to achieve. That is, the trouble is not in *conceiving* of a counterfactual in one's pessimistic thoughts are turned off.

[12] Markus Eronen, "Causal discovery and the problem of psychological interventions," *New Ideas in Psychology* 59, (2020): 100785.

One final, more intricate case. Suppose I am interested in the causal relevance of the stylization of some memo on individuals' reactions to the memo. I present the content memo in ordinary sentence case. I present its contents in All Caps. I capitalize every other letter; I randomly capitalize some letters and not others. And so on. In each case, I make sure to write the exact same letters on the memo, so my conclusions about the effect of the memo's stylization are not confounded by any semantic changes. That is, I intend to be intervening *only* on the capitalization pattern.

Now, suppose the memo I present my readers contains the following sentence:

(1) I find the polish to be nauseating.

and compare the reader's response to their reaction to the "same memo" when subject to an intervention wherein the letter 'p' in (1) is capitalized so that the sentence now reads

(2) I find the Polish to be nauseating.

No doubt, my intended intervention has failed, for the capitalization of 'p' in (1) *did* yield a semantic change to the word ('polish' to 'Polish') and in turn, to the sentence as a whole (from a harmless complaint to a rather inflammatory statement). In fact, no manipulation to the letter case of 'p' does not launder in these semantic changes.

The diversity of these cases complicates the seemingly tidy image of a surgical intervention making a change only to its target and leaving everything else "as is." But worryingly still is that among the troublesome variables and sets of variables are those that we in fact make plenty of use of in both our everyday and scientific practices of causal inquiry. These would seem to be the kinds of causal factors that interventionism would be particularly concerned with vindicating.

Let's review the cases and make a first pass at drawing out their features that are unruly for the interventionist. In the case of the gender non-conforming interviewee, the interventionist cannot

disentangle the causal factors of *sex status* and *gender conforming status* in a counterfactual that bears selective witness to a change in only one that leaves unperturbed the other. This problem generalizes. She faces the problem whenever she must contend with *extrinsic* or *relationally* defined causal factors. A manipulation that changes the value of any such extrinsic variable necessarily changes the state of at least one of those factors in relation to which the variable is defined. And if such factors may both have distinct causal relevance for the outcome in question, the interventionist cannot perform an unconfounded manipulation of just one of them. The interventionist psychologist faces a related issue when she considers a manipulation to an individual's *pessimistic thoughts* that does not bring in train any changes to their *feelings of sadness*. Here too there is a non-causal connection—though not as tight a conceptual one as in the preceding case on sex, presentation, and gender conformity—between the two causal factors that an intervention cannot prize apart.

The third case brings out something of a methodological lesson for interventionism. It is common for interventionist analysis to proceed at the level of variables and their values, without diving down to articulate what a given manipulation to the actual underlying causal factor represented by a variable in fact consists in. In the case of the memo, it is one thing to simply stipulate an intervention that changes only whether a letter is written in lowercase or in capital case and leaves everything else "the same." And at that level of abstraction, the intervention seems clear enough. But we saw there that the manipulation in question failed to constitute an intervention because the targeted change to only the sentence's orthography rang in too a change to its semantics that is causally significant to the outcome of interest, i.e., readers' responses to the memo. A tendency to stipulate an intervention-based change made to some variable's setting without considering what changes are in fact *realized* by the described manipulation can elide changes made to other parts of the scene's causal structure and thereby mislead causal inquiry.

Woodward himself warned against breezy assertions of interventions. Recall his suggestion that variables describe "quantities or properties for which there is a clear answer to the question of what would happen if they were to be manipulated or intervened on."[13] But while it is for the most part easy enough to consider a variable taking on a different value or to imagine some part of the world being different from how it is in fact, conceiving of an *interventionist counterfactual* requires more than just this. Causation and causal relationships are after all features of the world, not of models or of abstract mathematical objects. The interventionist must take care to think through what posited manipulations to some causal system entail, so to be sure that her theory can even apply to, let alone deliver the right verdicts on, those cases of causation that have largely been assumed to be eligible for interventionist analysis. Neglecting to spell out the details of what manipulations actually consist in—that is, failing to elaborate on the exact content of interventionist counterfactuals—has generated a reification fallacy that dogs much interventionist causal analysis. Claims that seem clear enough in variable talk (e.g., "intervene to change *only* whether the individual is having pessimistic thoughts and nothing else") are completely unclear when translated into counterfactuals.[14] This dissertation is concerned with one such class of cases that, I claim, present a significant challenge for interventionist causal analysis but that have failed to draw notice in part because of this tendency towards reification. In particular, quick substitution of talk of causal models and variables in place of the actual entities for which they stand has obscured the trouble that causal factors that are characterized *extrinsically* or *relationally* and those that bear non-causal relations to each other pose for

[13] Woodward, "The Problem of Variable Choice," 1054.

[14] Thus although interventionists often take it to be rather straightforward what interventions to variables and variable values correspond to, my claim is the matter is much more opaque than they have assumed. For example, Woodward writes in the Introduction of *Making Things Happen*: "Causal relationships, of course, have to do with patterns of dependence that hold in the world, rather than with relationships between numbers and other abstracta, but in the interest of avoiding cumbersome circumlocutions, I will often speak of causal relationships as obtaining between variables or their values, trusting that it is obvious enough how to sort out what is meant." My view is that perhaps laboring through such "cumbersome circumlocutions" may actually illuminate the shortcomings of applying interventionist analysis to many cases of causation. Woodward, *Making Things Happen*, 14.

the theory. This has led to what I take to be an undue confidence in interventionism's ability to provide an illuminating analysis of causation in the social world, wherein factors of this kind are everywhere.

In the chapters that follow, I put forth a challenge to an account of the causal structure of the social world limned by interventionism. My argument presses on the themes introduced in this chapter: the problem of variable choice and the notion of a surgical intervention that changes only its target and makes no direct changes to any other parts of the causal structure. It proceeds by presenting cases of causation in the social world that pose trouble for interventionism along these dimensions and then pulls back to diagnose why the social world might be replete with such causes that are so difficult for interventionists to wrangle. I have already here gestured at an important part of the answer. In the social world, things have causal significance not because of how they are in and of themselves but because of how they stand in relation to other things. To preview a case to come, how I react to being given orders by a fellow teammate on my swim team depends not just on facts that can be localized to the orders themselves—e.g., how loud the orders are; what the content of the orders are—it depends also on certain extrinsic qualities orders: whether, for example, they are legitimate according to the team's rules. Interventionism, I will argue, cannot easily absorb lessons like this into its paradigm.

The argument against an interventionist account of causation in the social world runs along the center median of this dissertation, but along its side I hope to convince the reader of several points that have a more distinctively normative flavor. Interventionism is important to contend with not only because it has continued to prove itself to be a worthy philosophical theory of causation, be it a metaphysical account of causation, or an account of causal epistemology, or a semantic account of causal statements, or whatever else. Even setting aside its merits on these grounds, it seems to me

incontrovertible that the theory successfully captures ideas about causation that occupy a central place in various of our actual practices involving causal theorizing. And so an intervention, as it were, into the details of interventionism the *theory* bears too on those *practices* that are greatly influenced by interventionist thinking. Two kinds are of particular concern to me: causal thinking in our scientific endeavors and in our ethical reasoning.

An approach to causal inquiry that takes as the gold standard of causal evidence methods that *isolate* causal effects along the lines required by interventionism has massively renovated experimental and quantitative approaches to causal inference in the social sciences. This transformation has been particularly controversial in research into the causal effects of salient social categories such as race and sex, which though not amenable to the setup of a controlled experiment, seem crucial to study from a perspective that explicitly looks to trace their causal impact on individuals' life outcomes. I argue in Chapter 2 that the interventionist ideal that undergirds much research in this area distorts our thinking about how race and sex act as causes. For it recommends that the primary aim of methods of causal inquiry is to extract the causal effect of *just* the category "itself" out of the effects of many features that correlate with the category. Sound inference underwritten by interventionist thinking will look to distinguish the causal effect of perceived sex status on interview outcomes from the causal effects of dress and presentation, the causal significance of race on police use of force from the effects of judgments of suspicion or dangerousness.

Many causal studies are indeed guided, in my view mistakenly, by these standards. And yet, I show that oftentimes, social scientific practice evinces a sort of wisdom about how these categories may have causal significance for various outcomes in the social world that extends beyond what interventionist theory lends. By walking through a case study of a social scientific experiment that is guided but not wholly held captive by the interventionist ideal, I set forth another challenge to

interventionism. The argument also leads toward an alternative interpretation of why such social scientific studies work as causal studies not beholden to the interventionist standard. They nevertheless work as causal studies, I suggest, because of how they *depart* from the interventionist ideal; how they pull not just on those considerations of causal relevance proper to an interventionist conception of causation, but on considerations that are distinctively *ethical* in nature. This insight points towards a better way to pursue causal inquiry about race and sex, which I propose in the final section of that chapter.

But while purging our causal thinking from any normative thinking about race and sex is an impossible venture, the interventionist ideal encourages us to try nevertheless. And as I have already mentioned, many social scientists take the task of separating out causal effects of the category of race or sex "itself" from its many potential confounders to be the most important step of any exercise of causal inference. This orientation within social scientific causal inquiry has had a spectacular and, in my view, utterly corrosive effect on our moral and political analyses in these areas. I argue in Chapter 2 that it for one, makes a social ontological error about what sex and race are as social statuses in our world. If many of the features that are correlated with race or sex are in fact *constitutive* of the category, then it is a mistake to probe the category's causal significance by disentangling the status "itself" from those social facts that constitute it. Once these categories have had these social facts stripped away, what remains is a causal factor that bears no resemblance to those markers that have immense causal and ethical import in our social world today. They are signifiers emptied of any signification at all. Why should our social sciences be concerned with the causal role of *these* categories to begin with? And why would we consider them with any special normative interest?

Indeed, as I will argue in the subsequent chapters, the interventionist ideal presses towards causal inquiry that reduces social kinds in precisely these ways. Unsurprisingly, the causal analyses

that result are ill-suited to illuminate the racial and gendered patterning of numerous social outcomes, and neither do they provide promising tools to diagnose harms and injustices along these axes and guide projects of racial and gender justice. Even worse, interventionist thinking sets forth an onerous standard of causal evidence to prove that categories such as race *are* causally significant at all to important social outcomes. A theory of race and sex causation so constrained limits the reach of key ethical concepts as well. In Chapter 3, I discuss its role in a dominant analysis of discrimination, an ethical concept, which been taken by many to be in part grounded in causal relations. When paired with a particular interventionist analysis of causation, discrimination looks to be a rare occurrence, and in turn, its normative significance rather minor. There is, I argue, plenty reason to be mistrustful of in such an account.

Chapter 1. Interventionism in the Social World

§1. Trouble for interventionism

Consider the following case.

SKIRT INTERVIEW

Billy, a male-presenting individual, is applying for a job and wears a skirt and facial makeup to his interview. After the interview, he is told that he has not made it to the next round.

When Billy wonders whether he was not advanced further along in the process because of his *sex* or because he is *gender non-conforming* in his presentation, he is asking after the causal goings-on of his interview experience. Which, if any, of these factors influenced the interviewer's decision?

If Billy is an interventionist, he pursues his query by considering two pairs of contrast cases. The first targets the Sex Status variable and compares what in fact happens to Billy's job prospects in SKIRT INTERVIEW (they are quashed) with what happens to Billy in a situation in which—to stick with the image of divine intervention—God swoops in to alter Billy's assumed sex, perturbing nothing else in the interview scene in the course of doing so, so that he is taken by the interviewer to be female as opposed to male. The second targets the Gender Conforming Status variable and compares SKIRT INTERVIEW as it is with a counterfactual scenario wherein God intervenes to change only whether Billy is gender conforming in his presentation; he goes from gender non-conforming to gender conforming. Each pair of contrasts is constructed to home in on one of the candidate causes of the interviewer's decision. Different interview outcomes across each pair of contrasts shows the causal relevance of the intervened-upon factor. If Billy's fate is the same across the actual and counterfactual cases, the factor is causally irrelevant to the outcome.

Already, a number of alarm bells may be going off. One question occurs immediately: what is it to manipulate Billy's assumed *sex status* in a way that makes no changes to any other causally relevant factors on scene? After all, if in the counterfactual, Billy is taken to be female, his wearing of

a skirt and facial makeup no longer constitute *gender non-conforming presentation*. And so it seems that a change made to his sex status that is accompanied by no further changes in what Billy wears and how he makes up his face nevertheless brings in tow another change: a change to his gender conformity. But this change would seem forbidden by the lights of interventionism, for its occurrence disqualifies the manipulation that targets Billy's sex status from constituting a proper intervention. How, then, can the interventionist analyze the goings-on in SKIRT INTERVIEW when causal factors are entangled in a way that defies the standard conceptualization of an intervention?[15]

Troubling questions abound for the interventionist who looks to illuminate a causal structure that features causal factors which are themselves non-causally related as are *sex*, *skirt*, and *gender conforming status* in SKIRT INTERVIEW. Elaborating them and seeing through to their answers will be the main effort of the rest of this chapter, but for now, I want to set them aside to move along to two apparently quite different cases.

SWIM CAPTAIN

> As captain of her high school swim team, Jal runs practices when the team's coach is absent. Today the coach is absent, and per the team's rules, Jal is charged with leading practice. She is giving orders to her teammates, and she senses that they are annoyed with the setup. Indeed, they are bitter and envious about her authority over them in running practice that day.

BAN THE BOX?[16]

[15] An interventionist might reply that a manipulation that changes Billy's sex status and thereby changes Billy's gender conforming status *does* qualify as an intervention, since the change to the Gender Conforming Status variable obtains by way of the targeted change to Sex Status. Hence, there is no problem with the fact that the values of two variables are toggled in one fell swoop. I address this defense more thoroughly in the following section, but for now, I will make two points. First, this is a non-standard interpretation of what constitutes an intervention. It is clear that on Woodward's account of interventionism, an intervention that targets X with respect to some outcome Y may only change the value of another variable Z if Z lies on a causal path that proceeds from X. But Gender Conforming Status is not *causally* related to Sex Status. The manipulation changes the two variables in one go. The interventionist's own framework makes this clear: notice that it is impossible to intervene to set the value of Gender Conforming Status to be independent of the variables Skirt and Sex Status. If the variables were indeed causally related, this should be possible. Second, this reply is unresponsive to the challenge that the case presents for the causal query at hand. The problem is that no manipulation to Sex Status is an *unconfounded* manipulation. And even worse, every change brings in train a change to Gender Conforming Status, which is precisely the factor whose causal relevance Billy is looking to disentangle from that of Sex Status.

[16] Studies have also shown that the benefits of Ban the Box policies are unevenly distributed. Researchers have noted a drop in the employment rates of Black and Hispanic men after such policies were passed and an increase in the racial gap in hiring callback rates. One hypothesis for these results is that employers are statistically discriminating against

> Faye is a manager at an accountancy firm that is looking to hire a new associate. Mathilde's application for the job made it through the first-pass check and has now come across Faye's desk. In reviewing the application, Faye sees that Mathilde has checked the box in the application indicating that she has a criminal record (commonly known as just the "Box"). She also notices a four-year gap in Mathilde's work history. After mulling it over, Faye decides that it will not be worth the debate likely to arise about hiring someone with a record to forward Mathilde's application through to the next round of reviews.

Are Jal's teammates annoyed because Jal is *captain* or because Jal is *giving orders* at practice? What was the causal relevance of the *checked box* on Mathilde's application indicating that she has a criminal history on Faye's decision to decline her candidacy? What about the *four-year gap* in her work history? Did Faye reject Mathilde because of the checked box on her application (indicating a criminal record)? Or was it the four-year gap in her work history that influenced Faye's final judgment? Or was it some combination of the two factors that caused the bad outcome?

As I will show shortly, standard interventionist analysis applied to SWIM CAPTAIN and BAN THE BOX? encounters trouble as well. And though at first glance there appears little in common between what goes awry in SKIRT INTERVIEW on the one hand and these two cases on the other, I will show that complications in the former—most notably, the trouble that the existence of non-causal dependencies among causal factors poses for the possibility of realizing an intervention—are afoot in the latter cases, too. The dilemma that emerges rather quickly and clearly in the case of the causal variables posed in SKIRT INTERVIEW will require more details and detours to unravel in the cases of SWIM CAPTAIN and BAN THE BOX? But in the end, I hope to convince that the same stumbling block lies at the root of the problem in all three. If my argument succeeds, these cases illustrate a problem for the interventionist that cannot be confined to a narrow set of tricky corner-

certain demographic groups, and lacking explicit information on criminal history, are nonetheless associating some applicants with having records. Jennifer L. Doleac and Benjamin Hansen, "The Unintended Consequences of 'Ban the Box': Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden," *Journal of Labor Economics* 38, no. 2 (2020): 321–374; Amanda Agan and Sonja Starr, "The Effect of Criminal Records on Access to Employment," *American Economic Review* 107, no. 5 (2017): 560–564. For evidence that Ban the Box policies do not harm racial minorities, see Dallan F. Flake, "Do Ban-the-Box Laws Really Work?," *Iowa Law Review* 104, (2019): 1079–1127.

cases. They suggest a deeper incongruity between the picture of causal structure offered up by interventionism and a good account of causation in the social world.

A predicament concerning variable construction is the thread that ties together the three cases. But while the problem that emerges shows itself at the variable selection stage of causal analysis, it indicates a much deeper worry for interventionism as a whole. It will therefore pay to return to the matter of variable construction before pressing on with the challenge presented by my cases.

I suggested in Chapter 1 that underestimating the significance and the difficulty of the problem of variable choice contributes to a troubling lack of clarity concerning the key question of when exactly a manipulation of a variable constitutes an intervention in the technical sense required by an interventionist theory of causation. I have already noted one reason for the relatively neglected status of this first modeling step of interventionist analysis. It seems to be a view broadly accepted among prominent interventionists that variable choice and modeling construction is ultimately a relative affair. In the eyes of these interventionists, interventionist verdicts are always model-relative, as all causal claims should be.[17]

But there is another reason for not belaboring the matter of variable selection, one which veers away from this more "let-a-thousand-flowers-bloom" stance toward causal modeling. Perhaps we need not wring our hands over the matter because there is a clear contender for a regimented account of variable selection that can fruitfully plug into interventionist analysis. That is, one might

[17] Model relativity is embraced to varying degrees in Joseph Y. Halpern and Judea Pearl, "Causes and Explanations: A Structural-Model Approach. Part I: Causes," *The British Journal of Philosophy of Science* 56, no. 4 (2005): 843–887; Joseph Y. Halpern and Christopher Hitchcock, "Actual Causation and the Art of Modeling," in *Causality, Probability, and Heuristics: A Tribute to Judea Pearl* (London: College Publications, 2010), 383–406; Christopher Hitchcock, "The Intransitivity of Causation Revealed in Equations and Graphs," *The Journal of Philosophy* 98, no. 6 (2001): 273–299; Christopher Hitchcock, "Prevention, Preemption, and the Principle of Sufficient Reason," *The Philosophical Review* 116, no. 4 (2007): 495–532; James Woodward, "The Problem of Variable Choice," *Synthese* 193, no. 4 (2016): 1047–1072.

take there to be little reason to exercise ourselves over the problem of variable selection because is more or less easy to resolve. One account in particular, proposed by Ned Hall and Tim Maudlin, has stood out as a promising approach.[18] In brief, that recipe goes like this: carve up the causal system under study into variables tracking the status of reasonably well-defined distinct spatiotemporal regions that interact with each other. Then, designate for each variable, value settings that correspond to the intrinsic physical character of that patch of spacetime.[19]

This way of defining variables plays well with the relational constraint imposed by the consistency requirement between variables and interventions. If distinct variables track conditions that obtain at distinct patches of spacetime, there in principle should be no reason why each variable in the system cannot successfully be the target of local intervention. We simply call on God to swoop in and alter the state of that well-defined region from this to that, leaving untouched the goings-on at any of the other regions of spacetime. Accepting this elaboration of variables, the interventionist can make good on her analysis of causal structure rather straightforwardly. She simply subjects each potential cause to a manipulation, which would seem to automatically be unconfounded, runs the clock forward, and sees whether the potential effect changes accordingly. Work remains, of course, to determine in each case what makes for the best way to carve up spacetime into variables—e.g., at what level of analysis should the variables be situated?, how fine and coarse-grained should the carving be?—but this account of variables paired with the overall interventionist test for causal relevance paints in broad strokes the outline of a promising reductive analysis of causal structure.

---

[18] Ned Hall, "Causation and the Aims of Inquiry," in *Statistics and Causality: Methods for Applied Empirical research*, eds. Alexander von Eye and Wolfgang Wiedermann (Wiley 2016): 3–30; Ned Hall, "Structural Equations and Causation," *Philosophical Studies* 132, no. 1 (2007): 109–136; Tim Maudlin, "A modest proposal concerning laws, counterfactuals, and explanations," in *The Metaphysics Within Physics*, ed. Tim Maudlin (Oxford: Oxford University Press, 2007), 5–49.

[19] While not endorsing the full recipe suggested by Hall and Maudlin, Thomas Blanchard and Jonathan Schaffer also note that variable "values allotted should represent intrinsic characterizations" in a list of "natural necessary conditions" of causal model aptness in their "Cause without Default," in *Making a Difference: Essays on the Philosophy of Causation*, eds. Helen Beebee, Christopher Hitchcock, and Huw Price (Oxford: Oxford University Press, 2017), 175–214, 182.

I want to start by showing that this approach to variable choice—call it the *Intrinsic Character Approach*—is ill-suited to the causal queries at issue in each of my cases. Its shortcomings foreshadow those features of causation in the social world that present a distinctive hurdle to causal modeling and interventionist analysis. To begin, consider a natural starting point for an interventionist analysis of the causal query at issue in BAN THE BOX?: a model that features the variables Checked Box and Four-Year Gap, tracking respectively whether Mathilde's application features a checked or unchecked box indicating a criminal record and whether Mathilde's work history has a four-year gap or no gap at all. Does this model accord with the recipe put forth by the Intrinsic Character Approach to variable selection? For starters, each variable certainly does refer to a localized physical feature of Mathilde's application. We can, as it were, point to the patch of space on her application where the criminal record box is checked and the patch of space that corresponds to the gap in her work history (let's suppose the job application form has a format that makes glaringly clear when an individual lacks work history).

It seems odd, however, to suggest that it is these physical markings that are potentially causally relevant to Faye's decision, for Checked Box and Four-Year Gap are not causally efficacious in virtue of these localized physical features. It is, rather, what these markings *signify* that causally matter to Faye's decision-making process. But these *meanings* are, of course, wholly unrelated to the physical character of the markings themselves. So, the Intrinsic Character Approach to variable construction seems to point to the wrong sort of thing as what is potentially doing the causal work in BAN THE BOX?.

Contrast this fault with what is off about carrying out the recipe to model the causal goings-on in SWIM CAPTAIN. A variable that represents Jal's captaincy lacks physical referent entirely. Here, we don't even know what the variable Captain points to. Where in the world is Jal's title as captain located? Instead, the variable tracks not how some localized bit of spacetime is but rather a social

status that in this case indicates some relational fact about the world. To have the status of captain in a swim team is to stand in a particular position of authority vis-à-vis a standard rank member of the team. There is no distinct patch of spacetime whose intrinsic properties correspond to Jal's being or not being captain.

The trouble that strikes an adherent of the Intrinsic Character recipe in SKIRT INTERVIEW is different still. If the causal variables Sex Status, Skirt, and Gender Conforming Status do correspond to the goings-on of particular regions of spacetime, they must track *overlapping* regions of spacetime. That Billy, a male-presenting individual, wears a skirt makes it the case that Billy presents as gender non-conforming. There are a number of ways to put this point, depending on one's ontological predilections. If you prefer events-talk, you might say that the event of Billy's wearing a skirt is *co-incident* with the event of his presenting as gender non-conforming. If you prefer property-talk, you might say that Billy's appearing at his interview instantiates two properties at once. The property of Billy's wearing a skirt and the property of Billy's presenting as gender non-conforming are *co-instantiated*. These details can be set aside, for my point is not committed to one or other substantive metaphysical view. The feature of SKIRT INTERVIEW that I want to highlight bears on them all. In assessing the causal significance of Billy's assumed sex status, his skirt, and the gender non-conformity of his presentation on his interview outcome, the interventionist will want distinct variables to represent each dimension of possible difference—and yet any variables constructed to track these features of the causal system will not correspond to *discrete* and *distinct* regions of spacetime as requires the Intrinsic Character Approach to variable selection.

If the causal questions posed in BAN THE BOX?, SWIM CAPTAIN, and SKIRT INTERVIEW are indeed eligible for our understanding, then before even proceeding with interventionist analysis, the difficulties encountered in attempting to apply this recipe to my cases teach a valuable lesson about variable selection and model construction. My cases of social causation show that the kinds of things

that are candidate causes in the social world ought not be defined and thereby constrained by how things *intrinsically* are, or what their *intrinsic properties* are like. In the social world, causes may be efficacious in virtue of their (non-localized) meanings, which bear no necessary relation to the (localized) physical markers upon which they attach (as in the case of the checked box in BAN THE BOX?); they can refer to slices of spacetime that overlap or even entirely coincide (as in the case of Billy's sex status, his skirt, and his gender non-conforming presentation in SKIRT INTERVIEW); or they may have no well-defined physical location at all (as in the case of Jal's captaincy in SWIM CAPTAIN). Variables constructed via a recipe that tracks only how things *intrinsically* are, or what their *intrinsic properties* are like will not do to serve as building blocks for an adequate analysis of social causation. The Intrinsic Character Approach to variable selection is ill-suited to the interventionist whose task it is to illuminate the causal structure of the social world.

***

Interventionists generally hold that their theory does not require endorsement of one particular accompanying theory of variables in order to give a valid analysis of causation.[20] A non-reductive approach to causal inquiry sees interventionism as providing an account of causal structure that is compatible with multiple different ways of modeling the causal goings-on of some system. On this view, there is no reason to be wed to a single theory of variable construction that will work well for all cases, let alone adopt a recipe for variable selection that takes in the total physical conditions of some causal system and outputs a single set of variables. On this front, interventionism is a flexible theory of causal structure. Even if the Intrinsic Character Approach to

[20] Those who take causal modeling to be more of an art rather than a science embrace a pluralism about models. See the preceding footnote for works that defend this rather free-spirited approach. James Woodward concedes that a causal system may admit multiple different causal models, all of which are apt, but also defends criteria for variable selection that make certain choices superior to others in "The Problem of Variable Choice."

variable selection makes for an inadequate analysis of the causal facts in one set of cases, some other way of carving up the causal system, it is thought, will be able to step up to take up the task.

I want to put pressure on this assumption. By showing a dilemma that emerges at the stage of variable selection to reveal deep problems for the core commitments of interventionism, I will argue that the central tenets of interventionism *cannot* be cleanly cleaved away from the Intrinsic Character Approach of variable selection and still stay intact as a theory of causal structure. The failure of the Intrinsic Character recipe to provide for an adequate theory of variable selection touches on something deep about how causation works in the social world that shows it to be incompatible with interventionism as a whole.

Standard interventionist analysis of SKIRT INTERVIEW offers an intuitive entry point into the problem. It will serve to repeat the alarm bells that are triggered along the way. Was Billy's assumed sex status a cause of his failure to proceed to the next round of interviews? The interventionist's pursuit of an answer looks toward a counterfactual contrast, call it SKIRT INTERVIEW*, in which *only* Billy's taken sex status is altered. In this interventionist counterfactual, Billy is taken to be female; his wearing a skirt and presentation as gender non-conforming both stay as is in the original case, untouched by the intervention targeting a change to Billy's sex status. So it must be that in SKIRT INTERVIEW*, Billy's (identical) presentation *does not* conform to what is considered gender normative for interviewees taken to be female.

Now let ring the alarm bells. How can this be? What exactly is this counterfactual state of affairs? It is important to get clear on what precisely has gone awry here. What is at issue in SKIRT INTERVIEW* is that, as described, the state of affairs is *internally incoherent*. There can be no such situation in which *only* Billy's assumed sex status has been altered, while no other causally relevant changes have been made. The conditions described in SKIRT INTERVIEW* simply cannot obtain.

That the variables Sex Status, Skirt, and Gender Conforming Status track conditions that obtain in *overlapping* patches of spacetime makes clear the contradiction inherent in realizing the unconfounded manipulation on which an interventionist theory of causal structure relies: in this case, a change that looks to disentangle the causal operation of Billy's assumed sex status, on the one hand, from that of what he wears and his gender presentation, on the other. So it seems that we can provide a good explanation for what stalls interventionist analysis of SKIRT INTERVIEW by appealing to the Intrinsic Character Approach to variable selection. The fact that the proposed variables do not correspond to *well-defined* and *distinct* regions of spacetime makes it clear why one cannot even *conceive* of a change to Sex Status that holds fixed Skirt and Gender Conforming Status. God's hand, even in its divine precision, cannot thread this needle. Spatiotemporal overlap makes it impossible to ring in a counterfactual that bears witness to the selective operation of mutually dependent but still distinct causes—counterfactual contrasts which, for the interventionist, are constitutive of causation.

It will take more work to see how the shortcomings of the Intrinsic Character Approach to variable construction in SWIM CAPTAIN and BAN THE BOX? prefigure a wholesale failure of interventionist analysis to illuminate the causal structure of these cases. To make headway, let us first elaborate their interventionist counterfactual contrasts.

SWIM CAPTAIN*

> The swim coach is absent at today's practice, and per the team's rules, the captain of the team is supposed to lead practice. Despite not being captain, Jal takes it upon herself to lead practice and gives orders all the same to the rest of her teammates. She senses that they are annoyed with the setup. Indeed, they are annoyed about what they take to be her illegitimate and presumptuous behavior in giving orders that are completely out-of-line.

BAN THE BOX?*

> Faye is a manager at an accountancy firm that is looking to hire a new associate. Mathilde's application for the job made it through the first-pass check and has now come across Faye's desk. In reviewing the application, Faye notices that Mathilde has a four-year gap in her

> work history. She does not see anything else in her application that might explain this hiatus in employment; for example, the box indicating a criminal record is unchecked. Faye concludes that Mathilde is one of those fair-weather workers who is ultimately uncommitted to their working life. She has no patience with these workers and declines to forward Mathilde's application through to the next round of reviews.

An interventionist should now have all she needs in SWIM CAPTAIN* and BAN THE BOX?* to issue verdicts on the causal relevance of Jal's captaincy on her teammates' annoyance and of the checked box on Mathilde's application on her employment prospects. After all, these contrasts tell what happens when all that is different is that Jal is no longer captain and Mathilde's box is no longer checked, which are precisely the counterfactuals that standard interventionist appeal to in figuring causal verdicts.

How do Jal's teammates feel when non-captain Jal gives the same orders as she does in SWIM CAPTAIN during practice? As it turns out, her teammates are just as annoyed, even more annoyed in fact, when non-captain Jal gives the same orders when coach is absent from practice ("The gall she has!"). With this counterfactual outcome, the interventionist rules that Jal's captaincy is either not causally relevant to her teammates' being annoyed at all, or might even be an attenuating causal factor in their levels of annoyance. In BAN THE BOX?*, Faye rejects Mathilde's application with even greater haste upon seeing an unexplained four-year gap in her work history ("We don't hire anyone who has a fickle commitment to their working life!"). The checked box on Mathilde's application indicating her criminal record is, the interventionist concludes, a *positive* factor in her candidacy for the job. But these causal verdicts are plainly not so. In SWIM CAPTAIN, Jal's teammates *are* annoyed because she has legitimate authority over them as captain of the team. So her status as captain *is* causally relevant to why they bristle at her while she gives orders during practice. Similarly in BAN THE BOX?, the checked box on Mathilde's application *does* give Faye pause and contributes to her ultimate decision not to move Mathilde along to the next interview stage. The interventionist simply gets the causal structure in these cases wrong.

What is at issue in these cases is not, as was the case in SKIRT INTERVIEW, that no conceivable intervention could selectively manipulate the value of one variable while leaving unperturbed the values of the rest of the variables in the system. The setups offered up in SWIM CAPTAIN* and BAN THE BOX?* are perfectly coherent. Still, it seems that these counterfactuals, despite their being well-defined, are not the right contrast cases for the purposes of assessing the causal relevance of Jal's captaincy to her teammates' annoyance or of Mathilde's criminal record on Faye's decision not to advance her candidacy. If the interventionist's causal verdicts are indeed wrong, where in the course of her analysis did she stray? To answer this question, let's retrace the interventionist's steps and return to the causal model and set of variables that generates these cases as the supposedly relevant interventionist counterfactuals in the first place.

A causal model that outputs SWIM CAPTAIN* as the right interventionist counterfactual implies that what it is to "hold fixed" Jal's orders is for the *intrinsic physical features* of her orders-giving to be the same across counterfactuals. In other words, the preceding analysis of SWIM CAPTAIN suggests that all that is required to intervene on the variable Captain and fix the variable Giving Orders at the same value is to ensure that the change does not alter anything about the character of Jal's orders-giving qua physical performance. That is, in SWIM CAPTAIN*, the volume of Jal's voice, her intonations, the utterances that make up her orders, where she stands as she gives the orders, and so on are exactly as they are in SWIM CAPTAIN. The causal model that underwrites SWIM CAPTAIN* represents Jal's orders-giving with a variable that tracks their intrinsic physical characteristics. A similar story applies to the set of variables that generate BAN THE BOX?*.[21]

But of course, as anyone who has ever given or received orders knows, there is more to the act of orders-giving than its vocal and bodily performance. Orders have other qualities to them that

[21] In BAN THE BOX?, the variables Checked Box and Four-Year Gap track the physical markings themselves: whether the box indicating a criminal record is checked or not, and whether Mathilde's work history has a four-year gap or no gap at all.

are invisible at the level of the physical world. Jal's orders at practice might be, for example, *authoritative* or *presumptuous*. They might be *legitimate*, or they might be *illegitimate*. These too are properties of her orders, albeit *extrinsic* ones. They are a matter not of how the orders are in and of themselves; rather they depend on facts that obtain "outside" of her orders-giving. In this case, they depend on facts about the swim team hierarchy and rules about who gives orders at practice when the coach is not around. And again, as anyone who has been on the receiving end of orders knows, these extrinsic features of orders-giving can causally matter too. Whether one responds deferentially or with resentment to being given orders might depend on these aspects of them. If this is so, then it is not at all obvious that what an interventionist *should* mean to achieve in constructing a counterfactual that "holds fixed" Jal's orders-giving is to ensure that Jal, as a physical matter, repeats the same orders-giving exercise to her teammates in SWIM CAPTAIN*. If there are extrinsic or relational properties of Jal's orders that may also be causally efficacious, then duplication of the orders' intrinsic properties does not suffice to ensure duplication of all that might be causally significant about them.

In taking the Intrinsic Character Approach to characterizing Jal's orders in variable form, the causal model behind SWIM CAPTAIN* neglects the distinctive causal roles that the *extrinsic properties* of her orders-giving might yet play. So, the change made to whether Jal is or is not captain appears, falsely, to bring along with it no changes to her orders-giving. After all, the orders *themselves* are the same across both cases; nothing prevents non-captain Jal from standing there making the same physical movements and giving the same vocal performance as before. But the change to her status of course does introduce new spurious causal structure, since it alters whether her orders at practice are *authoritative*. And although this change is concealed in the interventionist's implicit causal model,

it still matters causally-speaking for Jal's teammates reactions. For it is Jal's now *illegitimate* orders-giving that triggers her teammates' annoyance in SWIM CAPTAIN*.[22]

The existence of causally efficacious extrinsic features wedges the interventionist between a rock and a hard place. On the one hand, it is simply false that SWIM CAPTAIN* gives a counterfactual contrast to SWIM CAPTAIN in which all that is different that is causally significant to her teammates' reactions is Jal's status as captain. Her orders are also different; they are now no longer legitimate, and this difference matters causally-speaking for her teammates' reactions at practice. As such, this difference risks confounding her inquiry into the causal relevance of her status as captain. The standard interventionist test is led astray by the implicit set of variables that generated SWIM CAPTAIN* as the right interventionist counterfactual.

On the other hand, departing from an approach to variable selection that tracks the intrinsic character of various components of the system and admitting extrinsic properties into her causal model does her no better. For now she finds herself unable to realize interventions that can disentangle the causal relevance of an extrinsic property variable from that of another variable on which the extrinsic property variable non-causally depends. There is no intervention that strips Jal of her status as captain, while retaining the legitimacy of her orders[23]—just as there was no possible manipulation to be made that alters Billy's assumed sex status without changing his gender

[22] Causally efficacious extrinsic properties are at play in BAN THE BOX?, too. When the box indicating a criminal record is checked on Mathilde's résumé, the four-year gap in her work history does not trigger a double-take, for Faye takes it to be explained by the circumstances surrounding Mathilde's record. When the box is unchecked, the four-year gap is left unexplained, and Faye takes it to indicate Mathilde's weak commitment to employment. Hence, changing just whether the box is checked or unchecked changes whether the four-year gap is *explained* or *not explained*.

[23] What about a change that strips Jal of her status as captain but awards her legitimacy by some other means, say, by having the coach confer on her "orders-giving" privileges for the day? Would that manipulation qualify as an intervention on captain status, since it does not bring in train any change to the legitimacy of her orders? I think not, for such a change to the case would *add new* causal structure—in this counterfactual, Jal is not captain but has been nevertheless been picked out by the coach as having a special privilege at practice that day. This difference is likely to make for a causally significant difference vis-à-vis Jal's teammates' reactions, too. This response is a part of a general strategy to *control for* the "extra" causally significant changes rung into the counterfactual. I discuss this reply in greater detail in §3.2.2.

conforming status. The interventionist who concedes the inadequacy of the Intrinsic Character Approach to variable selection for figuring the causal structure of SWIM CAPTAIN finds herself back where she was with SKIRT INTERVIEW. Upon adopting a causal model that is in part comprised of variables defined extrinsically, she can no longer be sure to realize unconfounded manipulations that disentangle the operation of distinct causes. In such cases, interventionist analysis is dead on arrival.

The interventionist therefore faces a dilemma in my cases. She can either take on board the Intrinsic Character Approach to variable construction—and risk laundering in confounding causal relations that mislead inquiry into causal structure—or she can embrace a more liberal account of variables, allowing for the inclusion of extrinsic properties—but in so doing, since variable sets go hand-in-hand with possible interventions, risk impossibilizing the interventions that she needs in her analysis to disentangle the causal roles of different factors.

The trouble for the interventionist, however, is not just that a thing's being causally efficacious because of some extrinsic or relational feature poses a problem for the otherwise attractive Intrinsic Character Approach to variable selection. The deeper issue with extrinsic causes strikes at the notion of an intervention itself. They challenge the possibility of realizing an unconfounded manipulation of the target cause under study at all. If a thing's extrinsic property, by definition, non-causally depends on conditions "elsewhere," then changes that make a difference to those conditions can make a difference to how the extrinsic property is. When those conditions are themselves potentially causally relevant to the outcome of interest, no intervention can prize apart the two potential causes—the extrinsic property, on the one hand, and the conditions on which they non-causally depend, on the other—in a way that bears witness to the potentially distinct causal efficacy of each.

***

Extrinsic causal properties are an instance of a more general problem for interventionism. Whenever candidate causes that are considered distinct exhibit non-causally dependencies, the interventionist encounters trouble ringing in the counterfactual that disentangles the causal factors by selectively varying one and not any of its non-causally linked brethren. Insofar as comparison with this contrast case forms the basis of her analysis, unworkability of the unconfounded manipulation test is, I argue, crippling for interventionism's prospects for an adequate theory of causal structure.

Up to now, I have largely taken it to be uncontroversial that a thing's extrinsic properties *can* play a distinctive causal role in affecting some outcome in the social world. In §2, I defend this claim in greater depth and discuss why cases of non-causal dependencies among causal factors are so ubiquitous in the social world. Many causal relations in the social world, I contend, are structured by meanings and meaning-making. In my cases, what emerges as causally relevant to how an agent responds to some state of affairs, and how it is causally relevant, depends on how things are in the broader scene within which she is embedded. To return to my cases, how someone perceives a skirt in an interview context, reacts to being given orders, or judges a candidate's work history depends invariably on facts about the person who wears the skirt, the person who gives the orders, and the contents of the rest of the candidate's application respectively. Does the interviewee present as male or female? Does the individual who gives the orders hold a designation of legitimate authority? Does the applicant seem to have a reason for their absence from formal employment? A change made to these factors brings along with it, *simultaneously*, changes to the social meaning and so the causal significance of the former ones. This, I claim, is a general feature of human sense-making, and its incompatibility with interventionism's reliance on unconfounded manipulation to provide the key test for causal relevance poses a serious issue for the theory's capacity to illuminate causal structure in the social world.

I'll then go on to offer up and respond to potential interventionist resolutions to the challenge that my cases present in §3. One reply argues that the problem of non-causal dependencies among variables should be resolved by excluding from the start those sets of variables that fail to meet an *Independent Manipulability* criterion. Variables that do not meet this precondition, so the objection goes, must either be exchanged for a different set that can, or if no such variable set exists, the causal inquiry must itself be ill-formed. A second set of responses allows that interventionism can provide an adequate account of the causal structure in my cases but that the standard analysis must be revised in order to do so. From here, the interventionist has two options. She may either provide an explanation for why manipulations that are targeted to one cause but simultaneously ring in changes to other potentially causally relevant factors continue to be valid as interventions despite my claims to the contrary. Alternatively, she may outline how adjustments can be made to the standard interventionist analysis to adequately handle the challenge my cases pose. Of these two tacks, I will dwell much longer on the latter, which concedes that the preceding counterfactuals are misleading on account of their failing to present unconfounded contrast cases that isolate the causal operation of the variable under study. The proposed fix is to liberalize what constitutes an intervention on a variable by instituting modifications or *controls* to repair the counterfactuals so to retrieve the right causal conclusions, all while keeping in line with the spirit of interventionism.

In the course of replying to these objections in turn, I will end up retracing the steps of an argument that I have previewed in this section. I claimed that the interventionist who concedes the causal significance of extrinsic properties in SWIM CAPTAIN and BAN THE BOX? and looks to revise her underlying causal model accordingly meets the same dead-end that she encounters in her causal analysis of SKIRT INTERVIEW. My response to these interventionist replies will trace out an argument of the same form: Even while each of the interventionist's responses naturally aim at different cases—the objection based on the independent manipulability of variables more naturally targets

SKIRT INTERVIEW, while the response that looks to institute controls so to cancel out causal confounders triggered by imperfect intervention, speaks most directly to the difficulties afoot in SWIM CAPTAIN and BAN THE BOX?—I argue that the interventionist finds herself caught in a dilemma between them. In pursuing the latter reply strategy to rescue interventionist analyses of SWIM CAPTAIN and BAN THE BOX?, the interventionist is led to select an alternative set of variables that features the same trouble afoot in SKIRT INTERVIEW, thereby opening herself up to the rejoinders I put forth to the former.

In §4, I draw out from these responses general lessons for the task of variable construction and show that they cannot be both be heeded in the case of extrinsic causes. The problem, I conclude, stems not from the interventionist's making bad modeling choices, and instead arises due to the technical notion of an intervention that lies at the heart of her analysis. I close in §5 by returning to the key question of what it is about an interventionist theory that makes it unable to limn an illuminating causal structure of the social world.

§2. Intervening on extrinsic causes

Standard interventionist analyses of my cases either yield erroneous causal verdicts in the cases of SWIM CAPTAIN and BAN THE BOX? or cannot even get off the ground in the case of SKIRT INTERVIEW. Regarding the former two cases, it seems that something goes awry in the formulation of the interventionist's counterfactual contrasts, and here goes a preliminary diagnosis. In these cases, manipulation of the variable of causal interest, even if successful as an intervention that changes only the value of the target variable and leaves the values of all other variables in the causal model as is, realizes a counterfactual in which other features of the scene that are unaccounted for in the variable set have been altered in a causally relevant sense as well. The orders that Jal gives at swim practice *are* different when she gives them as team captain compared to when she gives them

as a standard rank team member—even while nothing about the orders *themselves* is altered by the change to her captain status. Similarly, the four-year gap in Mathilde's work history is perceived differently on her application when it is "explained" by some other feature of her candidacy (such as by a checked box indicating a criminal record) compared to when it is accompanied by no such explanatory factor. And so it seems that "interventions" on the variables Captain Status and Checked Box do *not* leave completely undisturbed the other candidate causes on scene. Aspects of Jal's orders and the gap in Mathilde's work history are changed as well in SWIM CAPTAIN* and BAN THE BOX?*, changes, which I contend are causally significant to the outcomes of interest, and thus whose absence in the interventionist's model and analysis risks misleading investigation of the causes under study.

The proposition that a thing's extrinsic properties have causal 'oomph' may be controversial as a claim about causation in the physical world, but in the cases of social causation presented here, I think it should be drastically less so. The causal query at issue in my three cases all concern how agents react to features of their social situation. My subjects have in common an attentiveness and responsiveness to aspects of their environment which focuses not on how things are like in themselves but rather how they stand in relation to other features of the contexts within which they are embedded. My claim is that such *extrinsic* properties have distinctive causal significance in these scenes.

The interventionist who analyzes SWIM CAPTAIN looks to distinguish a reaction to Jal's orders from a reaction to her position as captain of the swim team by constructing a counterfactual contrast case in which her status as captain is changed, while her orders are kept "the same." But she is hard-pressed to find such a state of affairs. If the authoritativeness and legitimacy of Jal's orders depend, non-causally, on the position she occupies within the swim team hierarchy, there is no reconfiguration of the scene that changes only that position while preserving the extrinsic property

of her orders' being authoritative and legitimate. No counterfactual can properly duplicate Jal's orders *in the causally relevant sense*, while changing the conditions that ground the effect they have on her teammates. To put it another way, the causal powers that her orders have are partly constituted by her elevated position within a group hierarchy—by her rank as *captain*. In SWIM CAPTAIN, her captaincy is partly what makes Jal's orders what they are qua causes of her teammates' reactions. That Jal's teammates recognize her in this role structures their perceptions of and thus responses to her orders. Their being issued by an authority is an extrinsic property of Jal's orders, and furthermore, it is in virtue of *this* property that her orders have the causal effect on her teammates' annoyance that they do.

Many commentators on interventionism have noted the dilemma posed by cases which involve causal factors that bear non-causal relations to each other.[24] Taking a detour to explore one such exemplary work by Alexander Prescott-Couch will serve to show where my challenge departs from and expands on previous ones, as well as foreshadow how it is that my variables-centric diagnosis of where interventionist goes wrong can unify seemingly disparate counterexamples to interventionism and thereby make for a deeper critique of the theory.

In "Causation and Manipulation," Prescott-Couch considers how a manipulationist causal inquirer would disentangle the causal relevance of the Episcopal church's ordination rules permitting female priests to a congregation's satisfaction with the Church from the causal relevance of the presence and actions of the female priest herself.[25] The manipulationist encounters trouble when she

[24] Most commentary has revolved around models with variables that are logically, definitionally, or conceptually related. Examples well-trod in interventionist discussion include the variables "saying 'hello'" and "saying 'hello' loudly"; variables for total cholesterol, low density cholesterol, and high density cholesterol; variables representing coarse-grained features of some drug and its finer-grained microscopic features; variables representing mental state/properties and the physical state/properties that the mental (non-reductively) supervenes on. Woodward responds to the problem such variables pose to variable selection and interventionism in "The Problem of Variable Choice" and "Interventionism and Causal Exclusion," *Philosophy and Phenomenological Research* 91, no. 2 (2015): 303–347.

[25] This case, called ORDINATION RULES AND PRIESTLY ACTION, appears in Alexander Prescott-Couch, "Explanation and Manipulation," *Noûs* 51, no. 3 (2017): 585–520.

looks to hold constant the congregation's interactions with the female priest while changing the Church rules which allow female priests. The problem is that changing the rules necessarily brings in tow changes to the *existence* of female priests. Hence, no intervention can be made at all on the ordination rules—even though those rules can certainly be causally relevant to the extent to which a congregation is satisfied with the Church. For Prescott-Couch, cases such as ORDINATION RULES AND PRIESTLY ACTION of what he calls "ontological dependence" spell trouble for the interventionist, as they show that "not every causal relation is a manipulability relation."[26]

A similar complication emerges in my cases. In SWIM CAPTAIN, a change to Jal's status as captain necessarily brings in tow changes to the authoritativeness of her orders, the property in virtue of which her orders have the causal effect they do. So no intervention can be made on Jal's status as captain—even though her having the rank of captain can certainly be causally relevant to her teammates' annoyance. In Prescott-Couch's case, the female priests are themselves extinguished when the Church's rules permitting female priests are extinguished, but in my case, a manipulation of one causal factor does not wholly eliminate the other. The factor's existence persists, but it changes qua cause. Stripping Jal of her captaincy does not make it impossible for her to still give orders to her teammates.[27] After all, she may still stand there issuing directives to her teammates, all the same. Might this difference in the extent to which one causal variable depends on another indicate that an interventionist may more easily overcome the challenge in SWIM CAPTAIN as compared to that in ORDINATION RULES AND PRIESTLY ACTION? I want to resist this thought and show that what goes awry in Prescott-Couch's case is more general than the label of "ontological

[26] Alexander Prescott-Couch, "Explanation and Manipulation," 485.
[27] Though one might ask whether directives given by someone who is not in a position of authority are still "orders." For example, if Jal were not captain and instead were a student at a different school who infiltrated swim team practice, it seems strange to call any commands she issues "orders" at all. The question of what constitutes an "order" further bolsters my point in this chapter that the question of what exactly a causal variable tracks and what it is to intervene in a way that leaves other variables undisturbed is a fundamental matter that has been seriously understudied in work on interventionism.

dependence" implies. This will in turn reveal the challenge to interventionism to be much more extensive than Prescott-Couch suggests.

Notice first that in ORDINATION RULES AND PRIESTLY ACTION, it is not the case that the *person* who is the female priest ceases to exist upon changing the Church's rules. Neither is it that she cannot perform the same actions qua physical movements. The key is that the person cannot occupy the same position and perform the same actions *in the way that causally matters*: as a priest, a person who is ordained by the Church as a religious leader and authorized by the institution to be someone who can perform religious rites. Hence, though I agree with Prescott-Couch that the existence of causal factors which are resistant to interventions—including "social properties" such as *being a female priest*—poses a challenge to interventionism,[28] framing the case as fundamentally about ontological dependence is misleading. Since the problem of "ontological dependency" refers to a particular model and choice of variables, such a diagnosis takes for granted the problem of variable selection and in so doing, obscures from view the wider scope of the challenge. That a variable representing *priestly action* fails to track anything at all upon making a change to the ordination rules is a problem with interventionist analysis when paired with a causal model of the case containing that particular variable. That is to say, while the charge of "ontological dependence" is apt with regards to this particular selection of causal variables, the root of the problem in ORDINATION RULES AND PRIESTLY ACTION does not depend on this choice of variables. It rather emerges from the causal dynamics of the case itself.

One way of seeing why the true challenge does not depend on how one carves the system up into variables is to consider an interventionist response that replaces the variable representing the female priest's existence and actions with one that represents the existence and actions of the person herself. With this new choice of variables, the interventionist no longer encounters the same

[28] Ibid., 503.

problem of ontological dependence. After all, the person herself does not cease to exist, and she might even still perform the same bodily actions! Yet reformulating the causal model in this way is clearly unresponsive to the challenge that Prescott-Couch poses. What matters in ORDINATION RULES AND PRIESTLY ACTION is that the person-in-the-position-of-priest's existence and actions *qua causes* is affected by the change made to the Church's ordination rules. Even while the person herself may persist after the rules have changed, her presence and actions function entirely differently as a potential cause of her congregation's happiness across the counterfactuals. No longer consecrated as a priest, the ex-priest might still be present at Church and perform the same physical actions before the congregation—only now, they are actions of a completely different meaning, perhaps of a person who seems to think she is a priest, or of someone who is playacting, perhaps even mocking, religious exercises. Concerns of *ontological* dependency set aside, still this is not the counterfactual contrast the interventionist has in mind. However one may choose to model the causal system, it is in virtue of the individual's status as a priest that her presence and actions are causally significant to the congregation's happiness. Any manipulation that alters this *extrinsic* aspect of her presence and actions threatens interventionist analysis. And so, like the cases that I have presented, Prescott-Couch's ORDINATION RULES AND PRIESTLY ACTION shows how causally efficacious extrinsic features are troublesome for the theory.

Thus although Prescott-Couch's case is indeed problematic for interventionist analysis, his diagnosis of why conflates a causal model constructed as a representation of some causal scene with the case's causal goings-on themselves. This mistake is ubiquitous in commentary on interventionism and often results in questions about the fundamental elements of interventionism—most notably, the variable construct and the idea of "possible" interventions—being blocked. Causal analysis that proceeds exclusively by reference to models can lead to a failure to notice how deep a problem with interventionism might run: whether, for example, a problem arises out of a bad and

unilluminating instance of model construction or whether a given case might pose a problem for the core idea of causation in interventionist analysis. Hence, it is commonplace to see challenges to interventionism that make a claim as to the latter met with defenses that counter in the vein of the former. But rejoinders to the challenge posed by extrinsic causes that look to get around the problem by overhauling the causal model and reselecting causal variables repeat the tendency towards reification and so are doomed. No way of carving up a causal system into variables can wrest causal factors that are extrinsic or relational in nature into a form amenable to interventionist analysis. They fundamentally resist attempts at unconfounded manipulation.

This is the core of the problem in SKIRT INTERVIEW. Billy's clothing is characterized by a distinctive set of physical characteristics—it is a skirt; it has a certain shape, color, pattern, texture and so on. It also, on Billy, has the extrinsic property of being *gender non-conforming*—an aspect of the clothing that speaks to how it is related to other features of the scene, in this case to the assumed sex of the person who wears it. Billy's attire can trigger in someone an adverse reaction in virtue of its having any one of these properties. Suppose I have an aversion to skirts. In my past I had a humiliating experience with skirts and because of that, I cannot stand to be around skirts or any bottoms that are not pants. And now Billy shows up to the lunch table donning his skirt. Seeing it triggers my bad memories, so I get up and move to eat my lunch elsewhere. In this case, I left the lunch table because of Billy's *skirt*. Just the same, Billy's clothing might be causally relevant to an adverse reaction in virtue of its *being gender non-conforming*. Contrast the preceding case with another. Suppose that this time I cannot stand people who do not dress in ways that conform to their assumed sex status. When Billy shows up to the lunch table, I see that he is wearing a skirt and leave the table because his clothing *does not conform to his gender*.

An interventionist who looks to construct a model that distinguishes the causal goings-on in these two cases must select a set of variables tracking both the *skirt* and *gender non-conforming* features

of the scene. The problem is that, having included it, the interventionist can no longer prize apart the distinct causal factors of *skirt* and *gender conforming status*. She cannot manipulate Billy's clothing from skirt to pants without changing, in the same go, whether Billy is gender non-conforming or conforming.

But just as the *gender conforming status* of Billy's clothing is an extrinsic cause, so is the *authoritativeness* of Jal's orders and the *justified-ness* of Mathilde's four-year gap in employment. Thus, if the obstacle to interventionist causal analysis of SKIRT INTERVIEW issues from the general problem that *extrinsic* properties pose for the theory, then little separates the interventionist's trouble in SKIRT INTERVIEW from her difficulties in SWIM CAPTAIN and BAN THE BOX?. My three seemingly disparate cases are thereby united in the challenge they pose for interventionism.

I have argued that extrinsic causes defy the task of variable selection. And lacking a solution, the interventionist cannot provide an adequate analysis of causation in systems featuring extrinsic causes, which includes many cases of social causation. She can, however, find a way out if she can revise her theory of variables so to accommodate the problem of extrinsic causes. Two avenues of doing so come to mind, which I offer on behalf of the interventionist. For one, she might propose adopting a condition that restricts variable selection so to exclude sets containing variables that cannot be the proper target of interventions. Such a constraint on eligible variable sets might also serve as a guide for choosing variable sets that *are* amenable to interventionist causal analysis, which in turn, would yield counterfactuals unblemished by the problematic confounding that trouble my cases. Alternatively, she might allow for imperfect interventions to ring in confounded counterfactuals but suggest ways of repairing them to recover dependencies constitutive of causal effects. Filling in the details of each of these strategies and showing how they cannot succeed in rescuing interventionism is the topic of the following section.

§3. Interventionist replies to the challenge

Since for most interventionists, the central test of unconfounded manipulation applies within the framework of causal modeling, many have suggested constraints on eligible causal models to ensure that all variables within may be targets of intervention. For these interventionists, an adequate analysis of causal structure requires that it be possible to target a change to the value of any given variable without simultaneously changing the values of any other variables on scene, lest the target variable's causal relevance be inextricably entangled with that of other variables. This means that all combinatorial settings of variable values in the system must in fact be possible. Only variable sets that meet this *Independent Manipulability* precondition are suitable for causal analysis.[29]

According to this strand of interventionism, what goes wrong in my cases can be traced back to the selection of variables that fail this condition. When the value of one variable constrains the range of values that other variables in the system can take on, each cannot, in all cases, be manipulated independently of the others and so cannot, in all cases, be subject to proper interventions. And without interventions, interventionism cannot construct the counterfactual contrast it needs to illuminate causal structure. The proponent of Independent Manipulability charges that cases with this feature do not so much pose a challenge for interventionism; rather, interventionism shows what's wrong with the causal query posed either in the case itself or in the model that is constructed of it.

[29] Woodward calls this condition "Independent Fixability" in "Interventionism and Causal Exclusion," 316; Brad Weslake advocates for an "Independent Manipulability" constraint on variable sets in "Exclusion Excluded," *International Journal for the Philosophy of Science*, forthcoming.

§3.1: The Independent Manipulability reply

SKIRT INTERVIEW best illustrates the kind of trouble for interventionism that the Independent Manipulability constraint on variable selection is formulated to avoid. Whether Billy is or is not gender conforming at his interview *just is* a matter of what his sex status is taken to be and what his dress and appearance are like. Gender conformity is a relational concept that relates assumed sex status and presentation with the predominant system of gender. When these causal factors are plugged in as variables in a causal model, the *non-causal dependency relations* among Sex Status, Skirt, and Gender Conforming Status impossibilizes the requisite unconfounded manipulations upon which interventionist analysis relies. It is impossible in such cases to ring in the counterfactuals that bear witness to the selective operation of mutually dependent causal variables—counterfactual contrasts which, for the interventionist, are constitutive of causation.[30]

The Independent Manipulability diagnosis of what goes wrong in (attempted) interventionist analysis of SKIRT INTERVIEW concedes that representing the causal dynamics of the case with this set of causal variables impossibilizes the manipulations required to prize apart the functioning of distinct causes but denies that any damning conclusions about interventionism follow. It is indeed the case, so this interventionist says, that a model that carves up the causal scene into variables that exhibit relations of, say, logical dependence, supervenience, or definitional dependence prevents

[30] That the presence of causal variables whose values are mutually dependent poses a challenge to the possibility of unconfounded manipulation is not wholly unfamiliar to the interventionist literature. The so-called "exclusion problem" concerns whether a theory of mental causation can rule mental properties as causally relevant to some outcome, even while they supervene on physical properties that are "sufficient" on their own as causes of the outcome. The apparent trouble that these cases make for interventionism derives too from the non-causal dependency among candidate causes. Any changes made to a variable corresponding to some mental property will require associated changes in the value of the variable for the physical properties on which the former supervenes. This tight connection means that the interventionist's test will be unable to disentangle the causal relevance of the mental from the physical. For any subsequent change in the outcome of interest could in fact have been due to the physical properties, which changed in tandem with the mental properties. And so, the *real* difference-maker could be the physical properties all the while. The mental ones are certainly along for the ride, but they themselves might yet be causally inert. Michael Baumgartner has challenged interventionism's ability to handle the exclusion problem in "Interventionist causal exclusion and non-reductive physicalism," *International Studies in the Philosophy of Science* 23, no. 2 (2009): 161–178; "Interventionism and epiphenomenalism," *Canadian Journal of Philosophy* 40, no. 3 (2010): 359–384; "Rendering Interventionism and Non-Reductive Physicalism Compatible," *dialectica* 67, no. 1 (2013): 1–27.

interventionist analysis of causal structure from getting off the ground. However, the reply goes, this shows no fault in *interventionism* as a theory of causation; it is rather a mark of a defect either in the causal model proposed for the case or in the causal inquiry that is posed in the first place. An objector in this vein thus claims that what's gone wrong in interventionist analysis of cases like SKIRT INTERVIEW should be laid down at the feet of the case itself or the set of variables chosen to represent it. Any causal inquiry or model that poses Sex Status, Skirt, and Gender Conforming Status as causal variables in a model is improper, ill-defined, or simply inscrutable from the perspective of causal understanding. It is in fact a virtue of interventionism that the theory allows us to detect such pathology.

One direction this response takes has only a negative upshot and so leads directly into a dead-end. If only those variable sets that meet the Independent Manipulability criterion—and by extension causal questions that can be mapped onto these kinds of variable sets—are well-defined and appropriate candidates for interventionist causal analysis, then interventionism simply cannot speak to the causal structure of many cases. As a proposed solution to the complications that arise when potential causal factors lack corresponding interventions, the Independent Manipulability condition certainly gets the job done—but only in the bluntest way possible, for the constraint simply eliminates cases that require such representations from causal consideration entirely.

I find neither pronouncing the case to be ill-defined nor pleading the causal fifth to be an attractive option for explaining away the challenge that SKIRT INTERVIEW poses. The query in the case seems to me not only to be a perfectly sensible to ask, but the causal difference at issue makes also for a normative difference. For one, there are important moral and perhaps legal differences between a case, call it GENDER REACTIONARY INTERVIEW, in which an interviewer is put off by Billy's skirt's being "women's clothing," another, call it ANTI-MALE INTERVIEW, in which an interviewer does not want to hire anyone who they take to be sexed male, and a third admittedly

more strange but still conceivable case, ANTI-SKIRT INTERVIEW, in which the interviewer is negatively attuned to the physical, say, "skirtness" qualities of Billy's skirt. It seems that telling apart these cases' causal structures is an important step to figuring these differences. An analysis that cannot meet the former task cannot set us up for the latter.[31]

An alternative, positive interpretation takes the Independent Manipulability criterion to be a guide for selecting variable sets that *will* be apt for causal analysis. That is, in place of a causal model that features variables which exhibit troubling non-causal dependence relations, she may construct another variable set that *is* congenial to standard analysis and proceed with business as usual. So, the interventionist who subscribes to the Independent Manipulability condition may yet have a positive response to my cases. If what is troublesome is not necessarily the causal inquiry posed in SKIRT INTERVIEW itself but rather the particular causal model that was constructed to represent it, then perhaps Independent Manipulability may redirect us toward more perspicacious ways of modeling causal structure that *are* amenable to interventionist understanding.

Whether this redirection tack can succeed depends on whether the causal system in question can be adequately analyzed without needing to represent it with variables that would fail the condition. If the causal structure of my cases can be properly analyzed without opting for variables whose values are mutually constraining, then the Independent Manipulability constraint poses no real threat to adequate causal analysis. While it culls away certain variable sets that are deemed impenetrable by interventionism, perhaps all such sets make for bad or unnecessary means of getting

[31] Disentangling such causal structures has been taken by many judges and legal scholars to be significant for debates about whether discrimination on the basis of gender presentation or sexual orientation constitutes discrimination on the basis of sex. In the *Bostock v. Clayton County* (2020) Supreme Court case, the majority opinion penned by Justice Gorsuch and the dissenting opinion by Justice Alito dispute what a counterfactual change made to an individual's sex status should "hold fixed" in order to properly assess its causal relevance. Robin Dembroff and Issa Kohler-Hausmann argue that the majority and Alito are locked in a stalemate because there is no non-normative answer about what it is to intervene to change "sex"—i.e., what must be held fixed, for the change to constitute an intervention—in order to assess discrimination in "Supreme Confusion About Causality At the Supreme Court," *CUNY Law Review* 25, no. 1 (2022): 57–92. I am sympathetic to this view and say more about what I take the relationship between causation and discrimination to be in the subsequent chapters.

at causal structure anyway. For what have become canonical exemplars of causation in the literature—cases of causation of a physical nature like the classic ones of Billy and Suzy throwing rocks at a window—requiring that variable sets pass the Independent Manipulability condition seems to present no onerous burden. It would seem to be quite natural to model these cases without resorting to variables that would fail the criterion, and doing so does not seem to require skipping over details in the story that might be relevant to its causal goings-on.

But can the same be said of cases of social causation? When Billy wonders whether his poor interview outcome was caused by his being taken to be *male* (rather than *female*), or his presenting as *gender non-conforming* (rather than as *gender conforming*), or his particular sartorial choice to wear this *skirt* (as opposed to *slacks*), he is precisely looking to distinguish the cases GENDER REACTIONARY INTERVIEW, ANTI-MALE INTERVIEW, and ANTI-SKIRT INTERVIEW. The interventionist can only approach these causal hypotheses by constructing a model with distinct variables that correspond to each of these dimensions of difference. There is no alternative set of variables to pivot to which satisfies the Independent Manipulability precondition that still makes progress on the causal query at hand. And so the interventionist who subscribes to the criterion is left with only the constraint's negative upshots. She either deems SKIRT INTERVIEW to itself be ill-formulated, or she confesses to having reached the end of her rope: the causal inquiry at hand, though well-defined, cannot be wrested into a form amenable to interventionist causal analysis.

This leads to the deeper reason why the Independent Manipulability precondition on eligible variable sets makes for an inadequate response to the challenge my cases raise for interventionism. Recall that the worry that motivates the condition is that certain non-causal connections among variables in a causal model impossibilize the unconfounded manipulations upon which interventionism relies. Requiring that variable sets meet an Independent Manipulation criterion, the thought goes, safeguards against the construction of models that prevent interventionist causal

analysis from getting off the ground. But a strategy that looks for different set of variables that can pass the bar of Independent Manipulability as a rejoinder to the challenge repeats the reification fallacy: it takes the problem raised by SKIRT INTERVIEW to derive from the particular set of variables chosen to represent the system rather than the underlying dynamics of the causal system itself. The problem is not that a particular *set of variables* impossiblizes interventions in the sense that the values that variables in a model take on can be toggled independently of each other. Rather, as I have argued, what matters for interventionism is whether the *causal roles* played by the various factors we are looking to distinguish in a system are independent of each other such that one can be manipulated without affecting the causal functioning of any of the others. In SKIRT INTERVIEW, it is this failure of independent manipulability of the actual causal factors themselves that is the problem.

When a particular way of carving up a causal system into variables is conflated with the system's underlying causal dynamics, even a causal model comprised of variables that meet the bar of independent manipulability may resist interventionist analysis. In these cases, interventions on variables may *still* introduce spurious causal structure that misleads inquiry. As I will go on to argue, this is in fact what explains what goes awry in the counterfactual contrasts proposed for SWIM CAPTAIN and BAN THE BOX?—despite the fact that the variables in each of their respective models *pass* the bar of Independent Manipulability. Independent manipulability of variables is therefore not a sufficient condition for ensuring that the intervention-based test yields sound causal analysis.[32] To demonstrate this point, allow me a detour that turns first to another objection, this time aimed at those cases.

[32] Neither does the satisfaction of the constraint guarantee that interventionist analysis will deliver the right causal verdicts. I give an example of such a case in §3.2.2.

§3.2. The No Controls and Controls replies

Consider a reply to my cases by a different interventionist who sets aside the austere constraint of Independent Manipulability. Instead of waving away my cases and causal models as inscrutable from an interventionist perspective, she takes them on board and looks to show how the core of her test of causal relevance can nevertheless be salvaged even when manipulations on target causes entail potentially causally significant changes to other variables and hence do not technically meet the bar of proper intervention.

There are two routes that such a revisionary account of interventionism might take. Each adopts a different approach to treating the "extra" potentially confounding changes that accompany targeted interventions. The *No Controls* interventionist insists that no extra efforts need to be undertaken to account for the changes at once rung into those factors that share non-causal dependencies with the target of intervention. Manipulations may not constitute surgical interventions, but if the changes rung into other causally relevant factors do not make for actual confounders of the effect under study, then they are benign and do not blemish causal verdicts issued in business-as-usual interventionist analysis. No harm, no foul. The *Controls* interventionist, by contrast, grants that changes to multiple variables may mislead inquiry by introducing spurious causal relations and effects that are not genuine features of the structure underlying the original system under study. On this view, rescuing the theory requires installing conditions in the comparator counterfactual to shore up a sound basis for the interventionist account of causation.

Prospects for the No Controls strategy hinges on successful defense of the claim that changes brought in tow in cases like mine *never* confound causal inquiry—a view that is, at least on its face, at odds with the core definition of an intervention as an operation that isolates effects of some cause of interest by leaving unperturbed all other causal factors connected to the outcome isolates effects of some cause of interest. The Controls response, on the other hand, concedes that

simultaneous changes may confound standard interventionist analysis. Her burden is to develop a principled account of how to institute "controls" to repair a counterfactual scene fraught with spurious causes, and then defend why this account that makes these departures to standard interventionist semantics nevertheless adheres to the core of interventionism.

§3.2.1. The No Controls reply

The No Controls interventionist denies that any modifications are needed to adjust for the changes simultaneously rung in by imperfect interventions. For the additional causal effects that permeate a system through non-causally related variables are simply not ones that can mislead causal inquiry. On this view, variables so connected may be permitted to move in tandem with changes to the target cause, as any such changes are non-confounders of the effect under study.

What would this reply mean for my cases? In the case of SKIRT INTERVIEW, it accepts that a change made to Billy's assumed sex entails a change to Billy's gender conforming status but asserts that this change cannot confound inquiry into the causal relevance of sex on the interview outcome. The change brought in tow to the variable Gender Conforming Status may be allowed to stand in the interventionist counterfactual contrast, no adjustments needed.

I, for one, find the declaration of no confounding patently implausible in the case SKIRT INTERVIEW. It seems to me intuitive that changes to whether Billy is gender conforming in his presentation *can* confound an inquiry into the causal relevance of his being taken to be male. A counterfactual scenario in which Billy is taken by the interviewer to be female and, since Billy's clothing and facial makeup are held fixed, gender-*conforming* and subsequently receives a good interview outcome cannot tell between GENDER REACTIONARY INTERVIEW and ANTI-MALE INTERVIEW. On the No Controls interventionist's approach, these different causal structures are observationally equivalent.

In the cases of SWIM CAPTAIN and BAN THE BOX?, insistence on non-confounding leads to issuing precisely the mistaken causal verdicts that make these cases challenges to interventionism. Recall their details from my introduction of the cases in the preceding section. The interventionist who draws on counterfactuals SWIM CAPTAIN* in which the swim team is even more annoyed at a non-captain orders-giving Jal and BAN THE BOX?*, in which Faye has even more misgivings about Mathilde's candidacy when she does not indicate a criminal record misidentifies the causal structure of the cases. She concludes that Jal's being captain and Mathilde's having a criminal history either play no causal role in or are even ameliorating factors of the swim team's annoyance and Faye's decision to decline to forward Mathilde's application. But these verdicts are plainly false. Jal's captaincy and Mathilde's checked box *are* causally significant in effecting the outcomes in the original cases—notwithstanding how the outcomes of interest might vary in some other counterfactual setups. The No Controls interventionist, who shrugs off these other changes as non-confounders, issues a reply to my three cases that gets their causal structure plainly wrong.

More broadly, the assertion that in cases like mine, changes to other causal factors brought in tow by a targeted manipulation simply cannot confound causal inquiry seems to me wholly unjustified to begin with. Such a claim might hold water when variables in a system are related definitionally or by supervenience relations, such that one variable is reducible to another and the underlying causal connections from each to the outcome are the same.[33] To use an example popular in the literature, a pigeon trained to peck at the sight of red pecks at the sight of a patch of scarlet *because* scarlet is a shade of red. What *causally matters* about that patch's being scarlet for the pigeon's

[33] Woodward takes the No Controls line in his defense of interventionism from mental exclusion-based arguments, which contend that the theory cannot account for the causal significance of mental properties. At the center of his argument is the claim that in targeting a given variable for intervention, one need not control for variables that bear non-causally dependency relations with the target variable. Woodward argues that when variables stand in definitional and supervenience relations such that manipulating one variable brings in train changes to another variable, it is not necessary to control for the latter to assess the causal relevance of the target variable, since it poses no risk of confounding the effect of interest. Woodward, "Interventionism and Causal Exclusion."

pecking behavior is that scarlet is a shade of red. So, the causal relevance of a visual patch's being scarlet, as opposed to its being turquoise, to the pigeon's pecking behavior is not confounded by its simultaneous change from being red to being not-red. Similarly, manipulation-based inquiry into the causal significance of patch's redness is not undermined by a simultaneous change as to whether it is scarlet. There is nothing distinct between the pigeon's responsiveness to the patch's being scarlet and its responsiveness to the patch's being red.

The same, however, cannot be said of the non-causally related variables at issue in my cases of social causation. The causal significance of Jal's *captaincy* and her *orders-giving* on her teammates' reactions are not reducible to each other nor is the interviewer's response to Billy's *gender conforming status* reducible to her response to his *skirt* and *assumed sex*. In each case, these are causal factors that might play distinct causal roles in influencing a teammate's feeling of annoyance or an interviewer's judgment of an interviewee's performance. The team's responsiveness to Jal's being captain is more than just a response to her giving orders and vice versa; the interviewer's response to Billy's presenting as gender non-conforming is more than just a response to assumed sex and the skirt that Billy wears. So, the distinct causal relations in my cases indicate that causal effects *can* be laundered in when multiple variables are changed at once. If that is right, there is no reason why causal confounding should not pose a live threat to causal judgments made without accounting for these changes. The No Controls interventionist's claim to the contrary is plainly unsupported.

§3.2.2. The Controls Reply

The Controls interventionist, on the other hand, concedes that a strategy of permitting non-causally related factors to vary in tandem with the intended target of intervention does risk confounding in my cases and therefore is problematic for the interventionist approach to causal theorizing. But, she maintains, the problem does not defeat interventionism, because it can be

remedied with an account of how the "extra" changes rung in to non-causally related variables can be "controlled for" so to protect against what would otherwise confound judgments of causal relevance. This tack looks to revise the manipulation-based test at the center of interventionist causal analysis. A successful approach will not only generate sensible and illuminating verdicts of causal relevance in my three cases, it will fill in the details of a general account for identifying the changes rung in alongside a targeted manipulation that make for spurious causal structure and thus call for controls.

It was this inability to keep an accurate accounting of spurious causal effects, claims the Controls interventionist, that led to the construction of the counterfactual contrasts SWIM CAPTAIN* and BAN THE BOX?* which in turn misled interventionist causal analysis. Interventions on variables that seemed naturally well-suited to serve as the basis of interventionist analysis of SWIM CAPTAIN—a variable recording whether or not Jal is captain and a variable recording whether or not Jal gives orders—in fact smuggle in effects that confound causal inquiry. This diagnosis illuminates a lesson for the interventionist's task of variable construction. The breakdown occurs because although Jal gives the same orders when she is not captain in that she speaks with the same intonation, the acoustics of her delivery are the same, and so on, her orders are not causally efficacious *in the same way* across the contrast cases. How her orders link up to her teammates' annoyance is substantially different causally-speaking across the contexts in which Jal is and is not captain—a fact elided by collapsing two different causal behaviors under a single variable setting. Thus even though interventionist counterfactuals constructed from these models *appear* to the modeler different only in the value of the target variable of inquiry—whether Jal is or is not captain; whether the box on Mathilde's application is checked or not—they are in fact different in other causally relevant ways.

By now the explanation for this causal variation is familiar. The orders that Jal gives when she is a standard rank team member take on the quality of *being illegitimate*—an aspect of her orders

that was lacking in the original case in which Jal is designated captain but which is now salient to her teammates and causally significant to their reactions. Though she gives the same orders, characterized intrinsically, across the contrast cases of interest, an extrinsic property of her orders changes. They go from being *authoritative* to *illegitimate*, and her teammates are responsive to this change. The same story applies in the case of BAN THE BOX?. Even while Mathilde's employment history remains identical regardless of whether a box indicating a past criminal record is checked on her application, what the gap in her work history means to Faye is substantially altered when the box is checked compared to when it is not. Its causal efficacy is in turn substantially altered across the counterfactuals. In the original case, the checked box on Mathilde's application explained away the four-year gap in her work history, such that without it, the "same" gap in BAN THE BOX?* takes on new significance for Faye. *What would lead Mathilde to voluntary exit the employment sphere?,* she wonders. The gap now matters differently to Faye's decision and in a way that is distinct from the fact that Mathilde lacks a criminal history.

The Controls interventionist therefore stands in agreement with the preliminary diagnosis I put forth in §2: the causal significance of extrinsic properties troubles the interventionist's ability to posit a counterfactual contrast that prizes apart the causal relevance of the target cause of interest from that of other causal factors in the system. Her proposed solution is to intervene, as it were, in these causal goings-on so to eliminate the spurious effects triggered by these associated changes, thereby repairing the counterfactual contrast and rescuing the core of interventionist analysis.

It is worth comparing this variant of interventionism, which concerns itself with the task of instituting proper controls, with the standard interventionist framework. For the Controls interventionist, the "right" counterfactual is not necessarily one in which God's nimble hands, with great poise, swoops into a scene to tweak a single variable before gracefully withdrawing without a trace. To the contrary, this strand of interventionism suggests that the spirit of the theory may in

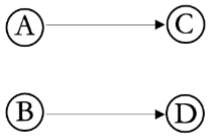
some cases be better portrayed by a flutter of activity, God's hands scrambling to make a suite of adjustments to various parts of the scene, so to construct the correct counterfactual contrast. Controls interventionism takes it that once we appreciate the fact that manipulations aimed at changing the value of one variable can entail other changes to the scene that are causally relevant, we are compelled to depart from the original picture of an intervention as consisting of a single surgically precise manipulation that directly changes the value of only the target variable. Better is a view of interventionism that focuses on the task of carefully composing a counterfactual contrast, which preserves as much as possible *causal sameness* across the two cases. Instituting controls amounts to making adjustments to a counterfactual setup in order to keep all else *causally relevant* the same. In elaborating an account of how to manipulate target causes while also clearing away confounding effects "accidentally" rung in, the Controls interventionist claims that her theory *can* rescue an adequate analysis of causal structure that holds fast to the spirit of interventionism.

The Controls interventionist's diagnosis that standard interventionist analysis may go awry in cases of failure to ensure causal sameness across counterfactual contrasts identifies a gap between an intervention that fixes all other *variable values* in a causal model and an intervention that holds fixed the *causal behaviors* of all other factors in the system under investigation. This in turn reveals a crucial ambiguity in the definition of an intervention. Namely, does the requirement that an intervention leave non-target causal variables undisturbed amount to a requirement along the lines of the former, concerning variables and variable values, or the latter, which concerns those entities that variables represent and their behaviors? [34] This question is crucial in causal analysis writ large, in both its

[34] One possible reason for the ambiguity between an intervention that holds fixed all other variables' *values* vs. an intervention that holds fixed the causal *behaviors* of all of the underlying causal factors that are represented by these variables may lie in the dual uses of the term "variable" to refer both to, in Woodward's words, "the properties, magnitudes, and so on related by causal claims and the representations we use to describe such properties." For, as I've argued, an underlying property can stay the same, and thus the corresponding variable can remain at the same value, and yet behave differently causally-speaking, i.e., a skirt may remain the same as a skirt but figure quite differently in the causal structure when worn by an individual taken to be male vs. one taken to be female. Woodward writes that he

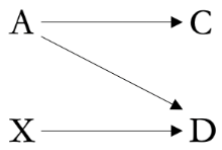
theoretical foundations and its methodological formalizations and operationalizations in scientific practice, and I will return to it throughout this and subsequent chapters. For now, it suffices to notice that the Controls reply brings to light once again the significance of the problem of variable construction for interventionism. If the failure to ensure causal sameness is indeed the culprit of what goes wrong in standard interventionist analysis of my cases, then the problem can be traced back to the set of variables that generated the counterfactual, which apparently do not capture all of the causally relevant factors in the system. This would explain how an intervention on a target that does not alter any of the factors represented by the other variables in the model can nevertheless ring in causally significant changes. These additional changes do not register in the causal model but generate spurious causal structure all the same. If interventionists are looking to realize a counterfactual contrast in which everything but the particular target cause under study is the same *causally speaking*, then ensuring only that all other variable *values* may be held the same upon intervening to change the value of the target variable will not do.

So, a factor can behave differently causally, even while the variable value that tracks some fact regarding the state that that factor of the causal system is in remains the same. A toy example based on a simple neuron diagram brings out well the general proposition. Consider a causal system comprised of the following neurons and their interactions. At time *t*, neurons A and B exhibit some set of firing behaviors, which then lead to some set of firing behaviors at a later time *t'* for neurons C and D.



The interventionist constructs the following causal diagram:

"believe[s] that the resulting conflation… [is] harmless," but in my cases where the two come apart, it does seem to generate problematic double-talk. Woodward, *Making Things Happen*, 377.



Binary variables A, C, and D track the behaviors of their identically named neurons in the causal system: value 0 corresponds to a *not-firing* behavior and 1, a *firing* behavior. The model represents the causal behavior of neuron B differently. The variable X takes on the value 0 if the neurons A and B exhibit the same firing behaviors, e.g., if neither fires or both fire, value 1 if neurons A and B exhibit different firing behavior, e.g., if one fires and the other does not. X's value in the model is therefore determined by a relational fact: how the firing behavior of neuron B relates to that of neuron A. In sum, variables are defined like so:

A: takes value 1 if neuron A fires, 0 if neuron A does not fire

C: takes value 1 if neuron C fires, 0 if neuron C does not fire

D: takes value 1 if neuron D fires, 0 if neuron D does not fire

X: takes value 1 if neurons A and B exhibit the same firing behaviors, 0 if neurons A and B exhibit different firing behaviors

The causal model consists of the preceding diagram paired with the variable definitions and set of structural equations, which indicate how variables C and D have their values updated according to the values of A and X:

$$C = A$$

$$D = 2AX - A - X + 1$$

It is straightforward to confirm that given information about the firing behaviors of neurons A and B, the causal model correctly recapitulates the firing behaviors of neurons C and D. What is more, variables A, X, C, and D defined in this way are perfectly independently manipulable. But when used to probe the underlying causal structure of the system, the model falters. For example,

suppose we want to know whether neuron A is causally relevant to neuron D. The method of analysis that interventionism recommends to answer the query is straightforward. Compare what happens to D in the following cases.

Case 1: A = 0, X = 0

Then according to the structural equations, C = 0 and D = 1. So, neuron D fires.

Suppose we intervene to change variable A's value from 0 to 1, holding fixed the value of X.

Case 2: A = 1, X = 0.

With these settings, C = 1 and D = 0. Neuron D does not fire.

The change made to variable A's value, holding all other variables at *t* fixed, *is* associated with a change in variable D's value at *t'*. So, the interventionist must rule that neuron A is causally relevant to neuron D's firing behavior, even while this is plainly false. No such interactions exist in the causal system.[35]

In the case of this simple neuron diagram example, a diagnosis and corresponding fix naturally suggest themselves. Intuitively, what goes wrong here can be traced back to the construction of the variable X to represent the behavior of neuron B. Since the value that X takes on does not straightforwardly map onto B's causal behavior, holding fixed the *value* of the former variable does not necessarily hold fixed the *behavior* of the latter neuron. In particular, while the value of X is held fixed across the two cases, the change to the value of variable A corresponding to a change to the firing behavior of neuron A brings in train, *non-causally*, a change to the firing behavior of neuron B. If X's value is to be held fixed at 0, the change to make neuron A fire is accompanied by a change such that B fires as well. B is therefore not "held fixed" across contexts, yielding the

[35] I thank Ned Hall for the suggestion to introduce such an example to introduce the point. A similar example and lesson appear in Section 3 of the extended version of his "Structural Equations and Causation." As he puts it there, the mistake in constructing the variable X in this way leads one to "conflate *logical* relations with *causal* relations." "Structural Equations and Causation," (2006), manuscript, 15.

mistaken causal verdict. Thus, what matters for interventionist analysis is whether the firing behavior of neuron B is held fixed upon manipulating to change the firing behavior of neuron A—whatever happens to variable X's value notwithstanding.

The lesson is this: constructing a causal model comprised of variables that are all independently manipulable does not suffice to ensure that changes to variable values correspond to independently manipulable *causal behaviors*. In the case of the preceding model, a change to the value of the variable A does not entail a change to the value of X; the two are wholly independently manipulable. But a change to A while holding fixed X *does* entail a change to the causal behavior of neuron B. Holding fixed X at 0, changing A from 0 to 1 brings in train a change in the causal behavior of B, from firing to not firing. In this case, manipulation of the variable A's value slips in new causal effects issuing from a hidden change in B's behavior that confound inquiry into its causal relevance on D—even while X is held to be the same in both cases. Therefore, if she wishes to conceive of interventions as manipulations that preserve causal sameness across counterfactuals, the Controls interventionist must attend to more than whether a given set of variables may be subject to independent manipulation. She must consider also how other parts of a system's underlying causal dynamics may shift upon making some targeted manipulation. This strand of interventionism brings to the fore the essential question of whether a given causal model and set of variables makes for a representation that risks misleading an interventionist analysis of causal structure.

That variables values in a causal model should track identical causal behaviors across relevant counterfactuals has not, to my knowledge, been explicitly articulated as a desideratum of variable construction. But it seems to me that something like it—which certainly concedes a non-reductionist analysis of causation—must lurk behind an analysis of causation that claims that an intervention on a given variable tracking some factor's behavior, which holds fixed all other variable values across counterfactuals, tells of that factor's causal significance. For if a single variable setting corresponds

to different causal behaviors across counterfactuals, the mere fact that a manipulation of a variable at *t* held fixed all other variables' values at that time would not suffice to show that the change in the counterfactual outcome tells of the causal role of the causal factor whose corresponding variable value was altered by manipulation. It could instead be that the change targeted at the factor of interest, say A, rang in a counterfactual state of affairs within which some factor B exhibits different causal behavior this time around—even while it is indeed not altered by the intervention and so its corresponding variable's value is preserved. In this event, the changed causal behavior of B in the underlying system could be the true difference-maker of a change in some outcome, rather than the causal factor A under study.

This suggests that what really underwrites the interventionist test is that each variable in a causal model whose value does not change corresponds to underlying factors in the system that too are not changed in their causal functioning across the counterfactuals. Only with this assumption in place can the interventionist make the claim that the manipulated variable corresponds to the only thing that is *causally* different and accordingly conclude that the counterfactual contrast case must therefore tell of *that factor's* causal role at time *t* and not of the changed causal operation of anything else on scene at that time. The upshot, then, is that variables in a causal model should be constructed such that each of a variable's values corresponds to a state that some part of the world could be in *that is causally efficacious in a similar way* across the relevant counterfactuals. That is, a well-constructed variable will take on values that correspond to some roughly uniform causal role that that feature of the world plays in the broader system, regardless of the values of other variables in the system. Call this approach to variable selection one that seeks *causal uniformity*.

Notice now something further: with the adoption of one additional and also seemingly rather minimal assumption that things that are intrinsically the same are causally the same, one arrives at the doorstep of the Intrinsic Character Approach to variable construction. So, one need

not endorse the proposition that there is something about localized intrinsic features *per se* that make for good variables in a causal model. All one must take on board is the thought that two things can only be causally different if they are different in their intrinsic physical states, and that if two things are duplicates of each other, they must display the same causal behaviors. Then, carving up a system into variables according to causal uniformity will result in variables that track the intrinsic character of various patches of spacetime. When a variable's causal behavior supervenes on its intrinsic physical state, the Intrinsic Character Approach to variable construction produces variables that are also causally uniform.

Renovating the causal model to accord with the causal uniformity criterion works to sort out the trouble that arises in the case of the earlier neuron diagram example. There, manipulating the value of variable A from 0 to 1, while holding fixed the value of X at 0, entailed a change to the causal behavior of neuron B. Setting X = 0 in one case corresponds to neuron B firing; in another, the same setting corresponds to neuron B not firing. The model fails to be causally uniform. Replacing X with a variable B with values that straightforwardly correspond to neuron B's firing behavior abides by the causal uniformity idea, and when joined by the structural equation D = B, yields a model that delivers the correct verdict that neuron A is not causally relevant to neuron D.

The Controls interventionist's claim is that what goes wrong in my cases echoes the mistake in the neuron diagram case. Just as a change to variable A that holds fixed the value of variable X entails a change to neuron B's firing behavior, in SWIM CAPTAIN, the change to Jal's captaincy, which holds fixed the variable representing whether she gives the same orders characterized physically, entails also a change to the causal efficacy of Jal's orders-giving. And this is a change to the scene's causal dynamics that takes place *in addition to* the change to Jal's captaincy status. Even while the variable representing Jal's orders-giving is "held fixed" at its same value—since Jal still performs the same physical act of orders-giving in SWIM CAPTAIN*—it matters to her teammates that her orders

are legitimate, and this aspect of her orders *is* altered when Jal is no longer captain. A causal model that does not account for the illegitimate nature of her orders when she is a standard rank team member in variable form launders in the (new) effect it has on the team's annoyance in a counterfactual contrast case that is supposed to have isolated the causal relevance of Jal's captaincy.

Can the same effort to construct variables along the lines of causal uniformity also restore interventionist analysis of my cases? A concern for causal uniformity proposes discarding the binary variable Giving Orders that records only whether Jal gives orders, for this variable represents the occurrence of an action whose causal effect substantially varies when transported to different situations. A more apt variable characterizes also the *kind* of orders that Jal gives: whether they are, for example, *authoritative* orders or *illegitimate* orders. Intuitively, this seems the right tack to take. Jal's teammates really are annoyed that she is giving illegitimate orders when she is not captain in SWIM CAPTAIN*.

But while the diagnosis that the variable Giving Orders' failure to be causally uniform is what triggers spurious causal structure befits the case, the proposed fix does not. Upon drawing up a causal model that includes a variable like Illegitimate Orders, the Controls interventionist now runs into the Independent Manipulability problem in earnest. With the new causally uniform model in place, it becomes impossible to manipulate variables independently of each other. No situation exists in which Jal is *not captain* but still gives *legitimate* or *authoritative* orders.[36]

As in the case of all variable sets which violate the Independent Manipulability condition, the reason for this impossibility boils down to the existence of non-causal dependency relations among the variables. What it is to be captain of the swim team is to have the standing that makes one's

[36] Or, rather, no situation exists that is different in no other ways that will confound inquiry into the causal significance of Jal's being captain on her teammates' annoyance in the original case. A situation in which Jal is not captain but her orders are nevertheless still legitimate is different from the case as described, in which the only way someone may give legitimate orders during practice when the coach is absent is for that person to be designated captain.

orders different from the "orders" of a member of the team who lacks the designation. There is thus a tight connection between *being captain* and *giving authoritative orders*. Jal's orders are authoritative *because* she is captain, where this "because" does not denote a causal connection. Rather, the authoritative nature of her orders is partly *grounded* in her status as captain. Modeling SWIM CAPTAIN with variables that adhere to causal uniformity thus recapitulates the same challenge posed by the selection of variables proffered for analysis of SKIRT INTERVIEW. There, recall, no manipulation was able to alter Billy's assumed sex status without changing too whether Billy presents as gender non-conforming. The dependency between the variables is conceptual. A person's gender conforming status depends (non-causally) on other facts about them: their assumed sex status and how they present—a matter which includes, among other aspects of social performance, whether they wear a skirt or slacks. Within our prevailing system of gender, to present as gender conforming while wearing a skirt, one must be taken to be sexed female rather than sexed male. And so no counterfactual can bear witness to the selective alteration of *just* Billy's assumed sex status to prize apart its causal relevance from that of the gender conformity of his presentation. In SWIM CAPTAIN, no intervention can change whether Jal is captain without changing also whether her orders are authoritative and legitimate. Part of what it is for one to be captain is for one's orders to fellow teammates to be authoritative and legitimate. Since what makes Jal's orders at swim practice authoritative as opposed to illegitimate does not inhere in how the orders are themselves, these are *extrinsic* aspects of her orders. They depend on how Jal and her orders-giving are situated within the broader context of the team's hierarchical organization, its rules about how practices are run, and on whether she is captain or a standard rank member of the team. If Jal's teammates are sensitive to this feature of her orders, then the causal significance of Jal's orders *non-causally* depends on these other factors of the causal system. Accordingly, no counterfactual can bear witness to the selective

alteration of *just* Jal's position as captain to prize apart its causal relevance from that of the legitimacy of her orders-giving.[37]

In SWIM CAPTAIN and BAN THE BOX?, what might confound inquiry are not additional distinct causal factors that may be cleanly struck out without affecting the functioning of other components of the system. Rather, the confounding features obtain in virtue of other facts about how the scene is set up. And because this dependence is not of the causal sort, these new factors are not candidates for being adjusted away with the right set of controls. A situation in which Jal is *captain* and *giving orders* just is a situation in which the orders that Jal gives are *legitimate*. The situation in which Faye sees that Mathilde has a *criminal record* and a *four-year gap* in her work history just is a situation in which Faye takes her employment gap to be *explained* by other facts about Mathilde's candidacy. These extrinsic features of Jal's orders and Mathilde's application depend non-causally on Jal's being captain and Mathilde's criminal record, but themselves have distinct causal relevance to the swim team's annoyance and to Faye's decision to pass on Mathilde's candidacy respectively.

This feature of non-causal dependency makes my cases notably different from others discussed in the literature in which what goes haywire in the interventionist counterfactual *can* be struck out with the right set of controls. Ned Hall presents the following as such a case. Suppose an interventionist is looking to tell the causal relevance of Billy's wearing an outrageously pink shirt on how his conversation with Suzy goes. The intervention that swoops in at time *t* during the conversation to change the color of the shirt while leaving everything else about the situation as it is certainly rings in the wrong kind of counterfactual. For in that counterfactual, Suzy takes notice of a paranormal experience and responds accordingly, exclaiming, "Your shirt just spontaneously changed color!" Hall diagnoses that in this case, the confounder that misleads causal inquiry *can* be eliminated if the interventionist makes suitable adjustments to other conditions so that the situation

[37] I do not walk through the case of BAN THE BOX?, though what I have said here applies there, too.

remains "normal." In the case of Billy's color-changing shirt, this means ensuring, among other things, that Suzy does not have a perceptual experience of his shirt that clashes with her immediate memory of it.[38]

The interventionist can successfully strike out causal confounders in cases like this because she can identify their effects as distinct from the effects of interest. Suzy's disorientation due to witnessing an apparently magical color-changing shirt is clearly not a response to Billy's wearing white instead of pink to lunch. So, it can be cleanly delineated as such and thus eliminated without threatening analysis of the causal factor of interest. By contrast, in my cases, the "confounding" effects cannot be struck out without changing other parts of the causal structure that do matter. Jal's orders being *illegitimate* cannot be struck out without changing her rank on the swim team; Mathilde's four-year gap being *unexplained* cannot be struck out without changing other facts about her candidacy that might be relevant for Faye's decision. Prescott-Couch describes the difference between cases like Hall's that can be resolved by appeal to installing "normal conditions" and cases like mine (and his), which cannot be, as coming down to whether "the combination of interventions *constitutes* rather than *causes* the abnormal situation." When an intervention on a part of the system *constitutes* an abnormality, efforts at instituting controls fail to restore the counterfactual to the right kind of contrast case; when the intervention *causes* an abnormality, controls aiming at restoring "normal" conditions are more promising.[39]

When causal factors share non-causal dependencies, they cannot be manipulated independently of each other. This, in turn, makes it impossible to appeal to some set of controls in order to reconstitute the "right" counterfactual contrast. So while the new, causally uniform models of SWIM CAPTAIN and BAN THE BOX? no longer hide away the sources of the spurious causal effects

[38] Hall, "Causation and the Aims of Inquiry," 25–26.
[39] Prescott-Couch, "Explanation and Manipulation," 498.

introduced by standard interventions, they land the Controls interventionist back among the host of challenges discussed in §3.1. When she cannot disentangle the causal significance of distinct causal factors, the interventionist faces two options, both in my view unattractive: conclude that the causal inquiry posed is ill-defined or concede that though the query is a valid one, interventionism cannot speak to it.

§4. Lessons for the art of causal modeling

My rejoinders to the Controls interventionist take on board her diagnosis that flawed variable construction misdirects interventionist causal analysis and proceeds to ask how one can renovate interventionism in light of this fact. Further challenges lie ahead for interventionism when trying to make good on this positive step.

To get to the worry, it helps to review the steps in the dialectic up to now. The Independent Manipulability and Controls replies each identify a distinct challenge to proper interventionist causal theorizing that issues from poor variable construction. First is a problem that blocks analysis entirely. When a model contains variables which are not independently manipulable, then it is not in all cases possible to intervene on a variable and set its value independently of the values of other variables. Insofar as the core of interventionism relies on such interventions to prize apart the causal relevance of variables, the analysis cannot even get off the ground. This is the problem that strikes the interventionist who looks to limn the causal structure of SKIRT INTERVIEW with variables Sex Status, Skirt, and Gender Conforming Status. The upshot of the challenge, however, is clear: given the risk of confounded manipulation, the interventionist must avoid variables sets that exhibit non-causal dependency relations with each other.

The second misstep in variable construction identified by the Controls reply does not stall causal analysis altogether but rather misleads it. When variables are constructed such that their

values do not track the causal behaviors of the factors they represent, the same variable value can correspond to a factor's having substantially different causal effects across different contexts. Models with these variables elide causally significant factors and thereby allow spurious effects to crop up in counterfactuals which seem to manipulate only the target variable but in fact bring in tow multiple changes to the system's causal dynamics. This is exemplified in the models that lurk behind the counterfactuals SWIM CAPTAIN* and BAN THE BOX?*. The proposed fix here is to construct models with variables that are *causally uniform* across contexts, ensuring that the model's variables represent all aspects of a scene to which the outcome of interest is causally sensitive.

These two morals seem to me to give reasonable constraints on variable selection. The hitch is that they cannot both be heeded at once. Adopting the perspective that variables should be causally uniform, the interventionist is led to construct models of my cases that track, for example, not just *whether* Jal gives orders—since a binary variable tracking Jal's orders *neat*, as it were, will have substantially different in the effect it has on her teammates' annoyance across contexts—but also the *character* of those orders. Are the orders that Jal gives during practice *authoritative* orders or *illegitimate* orders? A variable representing this more finely elaborated characterization of her orders plays a more stable causal role in different settings. And so intervening on such variables does not ring in counterfactuals that bring in tow spurious causes. But having reformulated the model of SWIM CAPTAIN in this way, the interventionist can now no longer make good on her other lesson: that variable values should not be tied together in a mutually constraining way that would prevent the ability to manipulate each independently of any others. One cannot manipulate Jal's captain status without bringing in train a change to whether her orders are authoritative or illegitimate. Nor is this a singular dilemma within which the interventionist finds herself. She encounters it any time she is pressed to represent causally significant extrinsic or relational facts in her model.

The causal uniformity criterion advises constructing variables that correspond to components of the system which do not exhibit much variation in their causal behaviors when transported across contexts. Since a model comprised of coarser variables necessarily means that the variables will track factors that exhibit greater variation in their causal behaviors in different situations, the condition pulls towards models that represent explicitly in variable form the relevant facts of some causal story at a finer-grained level of detail. In the social world, these causally relevant details often include extrinsic facts: how various components of the system are positioned in relation to each other. So, variables constructed to be causally uniform will sometimes need to incorporate extrinsic properties into their definition. On the other hand, the Independent Manipulability criterion seeks distinct variables whose settings do not impinge on the settings that any other variables in the system may take on. Models containing variables partly defined by extrinsic or non-relational facts violate this precondition.

The problem for interventionism, however, is not just that two reasonable criteria on variable construction cannot agree about what to do about an important class of causes in the social world, nor is it just that both routes lead to dissatisfying causal verdicts: either those that systematically mislead causal inquiry or those that cannot conclude anything at all. These are rather symptoms of a deeper problem, which becomes apparent at the stage of model construction but whose source lies in the conception of causation at the heart of interventionism.

I claimed in §1 that interventionism cannot claim to provide an adequate analysis of causal structure without confronting head-on the problem of variable selection. It cannot maintain, as its proponents often suggest, that the task of coming to an account of how to construct models apt for interventionist causal analysis may be separated from the core of the theory as a whole. The reason is that interventionism, as a distinctive counterfactual theory of causation, is distinguished by its account of what is to be altered by intervention and what is to be held fixed to yield the right

contrasts for which counterfactual dependencies are constitutive of causal relations. Variables, as interventionists' causal relata of choice, function as placeholders for the details of this account. This means that figuring principled theories of model construction and variable selection bears on the adequacy of the theory more broadly, for they fill in the content of interventionist counterfactuals. A set of causes that defies the task of variable selection entirely is therefore troubling for the theory's underlying counterfactual dependency account. If *no* choice of variables realizes the interventionist aim of changing only the target cause and keeping all else relevant to the system's dynamics the same causally-speaking, then there exists no contrast case that serves as the proper counterfactual, comparison with which reveals dependency relations constitutive of causal relations. In this case, interventionism is to blame, rather than the details of a given model. And insofar as it is a fact about causation in the social world that a thing's extrinsic properties can matter to how it acts as a cause, inability to wrest extrinsic causes into a form appropriate for interventionist causal modeling challenges the prospects for interventionism to offer a plausible analysis of social causation at all.

The challenge for interventionism posed by causally efficacious extrinsic properties easily passes notice, I think, because it is tempting to assume that there must be *some* way of constructing a model that can successfully illuminate the causal structure of cases like SWIM CAPTAIN and BAN THE BOX?*. Since the recipe for arriving at causal judgments proceeds rather straightforwardly once accompanied with a valid model, the first order of business for interventionist analysis is to figure a proper representation of the system, a good set of variables and causal arrows, that supports manipulations consistent with the right causal verdicts. When causal analysis goes wrong, interventionists will often return to the step of model construction to locate the source of the problem and propose a different model that will do the job.[40]

[40] For an example of this kind of sparring of models to critique or vindicate interventionism, see discussion of cases of late preemption and switching in Joseph Halpern and Judea Pearl. "Causes and Explanations: A Structural Model

This model renovation approach to settling whether interventionism makes for an adequate analysis of causal structure carries with it a notable risk, however, for it hazards a *reification fallacy*: model constructed as a representation of the causal goings-on of a system risks eclipsing in interventionist analysis what is in fact true about the system's causal interactions. Causal claims that may be true as a matter of how a model is constructed and the framework's accompanying set of assumptions stand in place of causal claims that are true as a matter of how causal reality is in fact.[41]

The tendency towards reification generates two worries for interventionism. The first concerns a distortion in the notion of *intervention* that can yield sound causal analysis. When the task of constructing the best model to represent the system's causal interactions takes center stage, it is easy to slip into an analysis of causal structure that revolves around a model-relative notion of intervention. On a model-relative conception of an intervention, the claim that a change made to some target variable held "everything else" fixed and thus passes muster as an intervention is a statement relativized to the set of variables that comprise the causal model. "Everything else" counts only those aspects of reality represented in variable form.

It is clear that the model-relative notion of an intervention cannot serve as the basis for an adequate counterfactual theory of causation. For what ultimately matters is not whether a manipulation to one variable's value alters the state of anything that happens to be represented by a variable in the particular causal model on deck. What really counts for proper causal analysis is whether the targeted manipulation directly changes other conditions that are *in fact* causally relevant

---

Approach. Part 1: Causes"; Ned Hall, "Structural Equations and Causation"; Joseph Halpern, "Appropriate Causal Models and the Stability of Causation," *The Review of Symbolic Logic* 9, no. 1 (2016): 76–102.

[41] Substituting talk about variables in for talk about those features of some scene that variables represent is often taken as a shortcut in the causal modeling literature. For example, Christopher Hitchcock writes: "[A]lthough the variables in a causal model represent various events that occur or might have occurred, and the equations represent patterns of counterfactual dependence among those events, it is often convenient to drop explicit talk of representation. Thus I will say such things as that A = 1 occurs, or that A takes the value 1 (rather than that the event represented by A = 1 occurs). Most of the time, this contraction will cause no confusion." The tendency of proponents of the framework skip over to engaging directly with variable settings and structural equations might partially explain common slippage into the reification fallacy. "Prevention, Preemption, and the Principle of Sufficient Reason," 503.

to the outcome of interest. These changes risk generating the very sort of confounding that threatens to invalidate the theory's central difference-making test—regardless of whether such features are explicitly represented in the causal model.[42]

The model-relativity concern is closely related to the second, I think more troubling, hazard of the reification fallacy. When models occupy such a central role in causal analysis, it is easy to find oneself in the position of simply declaring that a change made to some variable constitutes an intervention. Of course, to posit, with a nicely drawn model in hand, that a variable be manipulated in a way that rings in no other causally relevant changes to the system is easy. But whether there in fact exists a manipulation to some potential cause that realizes no other changes to causally significant factors in the system requires more than stipulation.

This brings to light an important distinction between how a manipulation is *described* and what a manipulation in fact *achieves* as a matter of altering causal interactions. Returning to SWIM CAPTAIN will serve to clarify it. Can we conceive of a manipulation that changes whether Jal is captain but preserves her giving orders at practice? Certainly. It requires no stretch of the imagination to picture Jal still standing there at swim practice, shouting out orders to her fellow teammates all the same, only this time, she is not captain. The problem is not that one cannot *imagine* such a scene. But to determine whether the only causally relevant difference across the two scenes is Jal's rank on the swim team, one must ask what that scene is a scene *of*. How does the change to her captain status alter what is happening in that counterfactual? Is it really a case in which upon

[42] To be fair, some interventionists are forthright in simply rejecting what I say here and take a model-relative notion to make for a perfectly adequate account of causation. I respectively disagree. When probing the causal structure of the world, it seems that we do not so much care whether X is causally significant for Y *in your model*. For the most part, the aim of our inquiry is to know whether X is causally significant for Y *in the actual world*. For a wholehearted endorsement of model-relative causal claims, see Joseph Halpern and Judea Pearl. "Causes and Explanations: A Structural Model Approach. Part 1: Causes." Christopher Hitchcock takes a more conservative tack by tacking onto his account, the notion of an "appropriate" model. X causes Y if it there exists an appropriate model, in which X is an actual cause of Y. It is unclear to me whether this analysis should also be considered to be model-relative. If there is disagreement on the matter of what constitutes an "appropriate" model, then it may well still be. See "The Intransitivity of Causation Revealed in Equations and Graphs."

intervention *all that is different* that Jal is giving orders as not-captain? This is a question more often obscured rather than clarified by abstraction.

My answer to it, well-trod at this point, is a forceful 'no.' To return to the image of divine intervention, when God swoops in to strip Jal of her status as captain, he realizes multiple causally significant changes to the scene at once. He makes it no longer apt for Jal's teammates to call her captain, for one. For another, he makes it the case that, given the coach's rules about who leads practice when she is not around and how different team members fit into the hierarchy structure (facts which are presumably held fixed across the counterfactuals), Jal is no longer the designated orders-giver on the team. So, God's change institutes a new regime in which Jal is not an authoritative source of orders; her orders are no longer legitimate. Notwithstanding the particular heading under which we might conceive of God's intervention—we might not *think* of God as swooping in to change anything about Jal's orders—a whole suite of changes is in fact realized by the move. The state of affairs rung in by a manipulation that makes Jal no longer captain is *at once* a state of affairs in which Jal's orders at practice are no longer legitimate.

When the interventionist assumes that there always exists a manipulation or set of manipulations that can target the right variables to finely tune which causal effects are turned on and off in a given contrast case, she confuses a choice of how one *models* changes to the goings-on of some situation with the changes that are *in fact* rung in by a given manipulation of the goings-on of the situation. This explains why the Controls interventionist couldn't repair what went awry in SWIM CAPTAIN*, no matter how diligently she revised her set of variables or precisely she targeted her manipulations. Given the swim team's internal rules and organizational structure, what it is for Jal to be captain is for her orders to enjoy a special legitimacy that orders given by non-captains lack. A change made to Jal's status as captain is therefore *at once* a change to the legitimacy of her orders, for whether her orders are legitimate depends (non-causally) on whether she is captain. The single

manipulation that strips Jal of her captaincy instantiates *at once* two upshots with distinct causal relevance: she is now a standard rank member of the team *and* the directives she now gives to her fellow teammates are now bossy and illegitimate. So long as Jal's status as captain is non-causally related to other things whose effects interventionist analysis look to distinguish and disentangle from the cause of interest, manipulation of her status as captain necessarily brings in train potentially confounding changes to these other features. What it is for members of the swim team to recognize Jal as captain is for them to recognize her orders as enjoying an authority that they lack when she loses the status of captain. No innovations in causal *modeling* can pull apart this dependency.

I have thus far argued in this chapter that *no* way of modeling my cases retrieves a promising interventionist analysis of their causal structure. No choice of variables and no ways of liberalizing the theory's manipulation-based test can spell out the content of a counterfactual that accords with the difference-making idea of causation at the heart of interventionism. This idea cashes out as a test that looks to *change* the target cause, while *holding fixed* all other causally relevant factors that are not the subject of causal inquiry. My claim is that maintaining this division is untenable when causes are extrinsic, for the state of the cause under study matters also to how other relevant factors are as causes. Changing it fails to hold fixed the causal roles of those factors that are not the target of analysis.

This leaves a choice. One can give up on the thought that how things are intrinsically does not exhaust all that is causally significant in the social world—deny, in other words, the causal significance of extrinsic properties. Or one can discard an account of causation defined by reference to a comparison of a pair of regimented counterfactual contrasts different *only* in the cause in question. To do the former and insist against what seems undeniably true about how the social

world works is, in my view, reckless. This leaves the judgment of the latter. An interventionist theory of social causation cannot hold.

§5. Theories of causal structure and practices of causal inquiry

Some critics of interventionism have pointed to a different defect of the theory's variable construct: that it suffers from an "excessive ecumenicism" which has the effect of authorizing too many variables as causes[43]—variables that by the lights of actual ordinary or scientific explanatory practices are simply not causes. This problem emerges because interventionists are, by and large, loath to place restrictions on the form that causal variables can take. And so the interventionist is left with a profusion of options for how to go about carving up the system into the causal relata and relations from which her analysis proceeds. The trouble, the charge goes, is not just that there are too many ways of going about variable construction but that many ways of doing so simply do not make for causal relata that are good building blocks of causal structure. Yet interventionists who eschew constraints on what a variable can be are unable to rule these out.

By contrast, the problem with variables that I have elaborated in this chapter is characterized by an opposite upshot. In my cases of social causation, in spite of this permissiveness in variable construction, no way of carving up the system pairs with interventionist analysis to yield the right causal verdicts. One upshot of this is that the theory cannot vindicate as causes many of the kinds of things that *are* taken to be causes in the social world in both common and scientific causal theorizing. But whereas the overinclusion problem arises out of an aversion to specify limits to interventionism's variable construct—though the generalized issue of figuring how to privilege the "right kind" of causal relata certainly dogs many other analyses of causation—the problem I have

[43] L. R. Franklin-Hall, "High-Level Explanation and the Interventionist's 'Variables Problem'," *British Journal of Philosophy of Science* 67, no. 2 (2016): 553–577, 556.

outlined here does not stem from the theory's conception of variables. It instead finds its source in the interventionist test itself. In tying a factor's status as a cause to the possibility of surgically intervening on its causal operation in a manner that holds fixed that of all other causally relevant factors, the interventionist bars herself from weighing in on the causal relevance of those factors that resist non-confounded manipulation. And so extrinsic properties or entities that are essentially extrinsically defined are excluded.

I want to suggest now that these two variable problems with their opposite consequences are two sides of the same coin. That is to say that what explains why the interventionist theory of causation is, on the one hand, unable to bear witness to the causal status of bona fide causes is the same as that which explains why interventionism, on the other hand, cannot circumscribe its set of certified causes to exclude those that are glaringly at odds with what our practices recognize to be causes. The diagnosis is this: even as interventionists take pride in having theorized an account of causation that hews closely to a scientific conception of causes and causal structure, the theory's formulations of the core constructs of *variables* and *interventions* are in fact impassive to scientific and ordinary practices of causal theorizing about the social world. In particular, the account of causation that is elaborated by interventionist analysis is developed orthogonally to the tried-and-true taxonomies for making sense of causation in our world, and it is this disconnect that generates the two variable problems.

The problem of excess ecumenicism follows from an account of variables that is insensitive to the scheme of explanatory kinds identified in our everyday engagement with and scientific inquiry about causes in the natural and social worlds. It is no surprise that such a theory of causation can pick out causal relata that are not fit to the kinds of things that are and are not causes by the lights of folk wisdom about causation and scientific study into it alike. The problem of extrinsic causes exposes a crucial assumption that lies behind an analysis of causation based on unconfounded

manipulation. Social science, which studies the causal behaviors of many extrinsically-defined explanatory factors and factors that are causal by virtue of their extrinsic properties, casts serious doubt on the interventionist's assumption that any cause must be conceptually possible to selectively manipulate in a manner that leaves all other relevant causes and causal relations in the system as they are.

But the premise that any cause can always in principle be disentangled from other causes in the system via surgical intervention has appeared to pose little issue for those cases of physical causation that have so far in the literature been taken as canonical cases of causation. From here, the success of interventionism at these scales is often taken to justify extending the theory to cases of causation in the social world. The thought is that if dependence under idealized unconfounded manipulations is a hallmark of causation at the level of physics and chemistry, then the same must be true of causation at those higher levels typically studied by the special sciences.

But as it turns out, our explanatory practices in scientific inquiry do not hew to many of these philosophical precommitments. Given the explanatory aims of social science, which studies (among other subjects) social relations and the meaning of signifiers and social practices, it is wholly unsurprising that many of the causal explanatory types that have emerged out of social scientific inquiry are essentially relational. Explanatory factors such as class and race and gender are, for most social scientists, defined not by reference to how an individual is like in and of themselves but by how they stand in relation to other things apart from themselves. Social scientific belief that these relational statuses are crucial in figuring the causal dynamics of many social phenomena is not called into question by the fact that few if any relational kinds feature in a good account of causal structure in the physical world.

The trouble with a philosophical analysis of causation that is insensitive to the kinds of causal explanatory factors that are picked out by our empirical investigations of causes and causal

relations is not just that it yields an overinclusive and underinclusive set of causal relata but that it is a bad approach to theorizing causal structure more broadly. Return to that classic representation of causal structure, of a diagram of nodes which represent causal relata and arrows which represent causal relations. To carve a system into variables is to snap a certain grid onto reality, grouping things together as a causal relatum represented by a variable node. Variations on that causal relatum correspond to different settings of that variable node. Interventionists who take a permissive approach to causal modeling allow that there are many such grids we can snap onto reality. Each of these grids affords a starting point from which analysis may proceed, drawing out the arrows that represent the causal connections between the grid's nodes, filling in the contours of the causal structure. Our folk causal wisdom and scientific inquiry endow us with many concepts and categories which may be explanatory causal relata, and so they give us candidate grids to snap onto reality.

One way of putting the problem with a theory of social causation advanced by interventionism is that it cannot limn a causal structure that interlocks with nodes on the kinds of grids offered up by our explanatory practices. In particular, if the interventionist starts with a grid comprised of nodes that track extrinsic factors picked out by folk wisdom and social scientific inquiry, proceeding with an analysis based on unconfounded manipulation will generate a structure of causal connections that does not reflect how these kinds are causally efficacious in our world. In the worst case, interventionism cannot vouch for the causal status of these extrinsic factors at all, since they cannot be targets of unconfounded manipulation to begin with.

The account of causal structure that results is one that cannot make sense of causal knowledge and reasoning we already have about the social world. What is at stake here is not just the prospect of vindicating "intuitions" but of coming to an account of social causation that fits with well-developed causal analyses carried out by careful scientific inquiry. An interventionist theory of

social causation that cannot do so is of little to use to any of our theoretical pursuits towards illuminating our social world or our practical pursuits toward making changes to and within it. These are the main, if not the only, standards that matter in evaluating the adequacy of causal analysis. What is more, they are the standards that matter *for interventionists* for whom it is considered a significant, if not the highest, virtue that their theory of causation accords with practices of scientific inquiry. An adequate account of causation is not only compatible in the sense of not being at odds with our best scientific practices but can rationalize existing practices and even guide future ones.[44] For an interventionist like Woodward, a good philosophical account of causation should pay close heed to the theories and "methods for investigating particular scientific domains" which in turn "should be attuned to the entities and structures those domains contain." A compelling analysis of causation is distinguished by its "methodological fruitfulness," which we judge by looking to the "actual practice of those involved in causal reasoning in various domains."[45]

My aim in this chapter has been to show that as a matter of these criteria that interventionists themselves hold in high esteem, their theory does not make the grade in the case of the causal structure of the social world. It is in large part due to the theory's failure to make room for knowledge gleaned via our everyday experiences navigating the social world and our scientific investigations of it, suggesting that interventionism's aspiration to elaborate a theory of causation

[44] See e.g., in James Woodward and Christopher Hitchcock's words: "[W]e take it to be an advantage of our approach that it makes clear the connection between counterfactuals and the sorts of manipulations actually carried out in experiments used to test causal and explanatory claims." "Explanatory Generalizations, Part I: A Counterfactual Account," *Noûs* 37, no. 1 (2003): 1–24, 9. Woodward explicitly construes of interventionism as "a set of methodological proposals, rather than as a set of theses about the ontology or metaphysics of causation," where he takes questions of methodology to be centrally concerned with "how we ought… to go about investigating, learning, and reasoning about various aspects of nature, about what sorts of theories we should construct, and about how we should reason about various important concepts in the scientific enterprise (such as 'cause')." "Methodology, Ontology, and Interventionism," 3577, 3588.

[45] Ibid., 3597.

that coheres with our best causal practices is on target.[46] All the greater shame, then, that it remains unactualized.

[46] Woodward foresees a potential need to revise one's theory and methods in light of the particularities of a given domain. As he puts it: "[D]ifferent sorts of investigative and reasoning methods may be fruitful for different sorts of entities and structures, depending on the features of the latter. For example, if the correct cognitive neuro-ontology is that the basic structures or units of analysis in the brain are distributed networks of various sorts, then different methods for identifying and reasoning about these will be appropriate than if one thinks that the basic units are highly localized neural areas. As another example, if the ontology$_1$ of some domain is that it contains structures in which values of key variables change over time in a way that is causally influenced by previous values of those variables and complex feedback relationships are present, generating data in the form of time series, such domains will likely require different methods of causal analysis than structures which are acyclic and can be assumed to have settled into some sort of equilibrium state which generates cross sectional data." I agree wholeheartedly with what Woodward writes here. And one way of reading my central argument in this chapter is as a claim that puts forth that the particular entities and structures of the social world call for perhaps not just a revision to the particular methods of causal analysis but an overhauling of the interventionist theory of causation as a whole. Ibid., 3580.

Chapter 2. Interventionist Social Causes: The Case of Race and Sex

§1. Introduction

In the preceding chapter, I put forth a challenge for an interventionist theory of causation that centered on cases featuring causal factors that are *extrinsic* in nature. For an extrinsic cause, what matters causally-speaking is not how the causal factor is in and of itself but rather how it stands in relation to other goings-on in the system. Extrinsic causes, I have argued, are trouble for interventionism, for their presence in some causal scene can impossibilize the prospect of surgically intervening on, or achieving an unconfounded manipulation of, the target variable of causal interest. With no possibility of intervention, interventionist causal analysis quickly meets a dead end. So, if the causal structure of the social world is dotted through with causally efficacious extrinsic properties, interventionism will be ill-equipped to provide an adequate account of it.

In this chapter, I bring philosophical theorizing about social causation into contact with real empirical analysis of the causal structure of the social world. My focus will be on the categories of *sex* and *race*, two of the most highly studied explanatory kinds in the social sciences and whose causal capacities in a social system exemplify the extrinsic causes dilemma for interventionism. Probing modern social scientific methods that aim at identifying causal effects of race and sex and clarifying the interventionist-like logic that undergirds them will serve to show the cash value of the argument presented in Chapter 1, extend that critique of interventionism, and also guide us positively towards better philosophical analysis and social scientific inquiry into the causal structure of the social world.

Social scientific approaches to studying the causal effects of race and sex are many and varied, but all set their sights on the same horizon goal: the idealized controlled experiment. In the idealized controlled experiment, two contrasting setups are different with respect to the potential cause under study but are different in no other ways. Then, the logic goes, having eliminated all

other causal factors as potential difference-makers to the outcome of interest, any difference in the outcome that is observed in each setup gives the causal effect of the only factor that was different: the potential cause under study.

What underlies the diverse kit of causal methods is the thought that the hallmark of causation is a difference-making relation that obtains in highly controlled conditions of experimentation—an idea about causation that sits also at the core of an interventionist analysis of causation. This shared foundational view on causation might now seem to contradict a conclusion I drew in the final section of the previous chapter: that it is in part because interventionism is *insensitive* to the practices of social scientific causal inquiry that explains why it cannot account for extrinsic causes and so fails to provide a plausible analysis of causation in the social world. But how can this claim be true if it is the case that, as I have just now admitted, social scientific causal analysis of race and sex *is* modeled on interventionist thinking about what it is for the category to be a cause? It would seem then that social scientific and interventionist causal analyses of race and sex either stand or fall, together.

While predominant causal methods in the social sciences, by and large, do take idealized manipulations under all-else-equal conditions to be the gold standard of causal inquiry, I want to show in this chapter that good careful causal inquiry into race and sex in fact *deviates* from an analysis that would follow from a naïve operationalization of interventionist causal thinking.[47] And looking to where social scientists depart from the strictures of the idealized controlled experiment—setting aside the matter of whether they do so consciously or with some guiding background theory in mind—reveals something wholly new and surprising about causation in the social world. My claim here is of a piece with a point I made in the previous chapter: that good social science evinces a

[47] I add in the modifier "naïve" here not because I think that a more sophisticated operationalization of interventionism would get causal inquiry right, but because, as I will go on to show, it's not clear *what* interventionists *would* recommend is the right operationalization of their theory in the case of figuring race and sex as causes.

wisdom about social causes that can and should inform our construction of metaphysical theories of causation in the social world.

At the center of my argument in this chapter is a puzzle about *audit studies*, which I present in §2. Audit studies, and their closely related correspondence studies, are a type of field experiment that mimic the conditions of idealized controlled (or randomized) experiments and test for the causal effect that some characteristic of an individual has on decision-making. Such studies might, for example, randomize Black-sounding and white-sounding names onto fictitious résumés, send them out to employers looking for hire, and observe the rates at which Jamals and Gregs receive callbacks, interpreting any disparity between them as giving the causal effect of race on callback outcomes. The prevailing consensus takes audit studies to present strong evidence for causal effects because of their elegant controlled experiment-like design. The Greg and Jamal in a résumé correspondence study are, after all, "identical" applicants but for their (raced) names. So it is typically thought that what is most compelling about audit studies is their ability to wrangle social statuses whose causal effects are particularly tricky to study in "controlled" settings into a design that approximates the gold standard of causal inquiry.[48] By contrast, I hope to show in §3 that the design of audit studies reveals the significance of forms of reasoning that are *unaccounted* for—or to put it more starkly, are strictly *barred* from entering—in the setup of an idealized controlled experiment. Attention to when an audit study that probes the causal effects of race or sex succeeds and even more importantly, how such a study can fail, reveals the ineliminable role that substantive moral and political reasoning plays in figuring what it is for race and sex to be causes. If this is correct, then it is a mistake to try to

[48] "[A]udit studies manipulate the second part (race) to directly capture the first part (differential treatment) of the definition. Thus, by carefully controlling and counterbalancing all other variables in the experimental process, audit studies provide strong causal evidence of discrimination." S. Michael Gaddis, "An Introduction to Audit Studies in the Social Sciences," in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. M. Gaddis, (Springer, 2017), 9.

separate out our causal theorizing about race and sex and our normative theorizing about things like racism and misogyny. This view on causal structure departs from the standard line taken by all leading theories and theorists of causation. I offer the interventionist room for reply in §4 and respond to an objection that posits that ethical theorizing can be stripped out of the causal theorizing.

Interventionism's inability to respond to the puzzle presented by audit studies is a weakness descending from its trouble with extrinsic causes, but it foreshadows still more worrying deficiencies ahead. In §5 of this chapter, I draw out these shortcomings by returning to the theory's central commitment to illuminating a factor's causal significance by disentangling its effect from that of all other distinct causal factors. Most social scientists and theorists who study stratifying social categories contend that a crucial part of how they structure our social world lies in the roles that the innumerable things which systematically correlate with those categories play in the broader social system. A theory that defines causal effects of race and sex by dissociating them from other goings-on stands in direct opposition to this approach and is ill-suited to show how such regularities may yet be a part of a causal accounting of race and of sex themselves, and so a part of a normative accounting of racial and gender injustices. If so, the interventionist loses her good standing not only among social scientists but among metaphysicians. For it is the province of both to have concern for the explanatory and practical deficiencies that their concepts and categories might yield. A better way forward starts from the insight that causal theorizing about the social world cannot do without normativity. In §6, I return to salvage audit studies as causal studies despite their failure to conform to the interventionist ideal and suggest that it is these studies' implicit ethical claims that in fact aid their efforts to make causal claims. I close in §7 by reconfiguring metaphysical lessons for interventionism as a methodological view of causation.

§2. A puzzle about audit studies

Over a span of several months in 2001 and 2002, economists Marianne Bertrand and Sendhil Mullainathan sent out fictitious résumés to employers looking for hire in Boston and Chicago.[49] Each employer that the researchers contacted received two "matched pair" résumés, making for a total of four résumés, from a bank of résumé templates that Bertrand and Mullainathan had created. Each pair was made up of a "Black" candidate and a "white" candidate with résumés that were "matched" or very similar in their contents. And so Greg and Jamal applied to jobs with résumés which were "identical" but for race. With this setup, the logic of the study goes, the difference between the rate at which Jamal received callbacks from employers and the rate at which Greg received them gives the causal effect of *being perceived as Black versus as white* on callback outcomes. After all, if the only thing that differs across the two résumés is the applicant's race signifier, the difference in the callback rate must tell the causal effect of race on employers' decision-making. Nothing else about the two candidates could have been the difference-maker.[50]

---

[49] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal: A Field Experiment on Labor Market Discrimination," *American Economic Review* 94, no. 4 (2004): 991–1013.

[50] There has been dispute about whether the names that Bertrand and Mullainathan chose in their study are indeed different only in the signal of race and do not also differ in their signal of, say, class or socioeconomic status. That these latter factors might have confounded the original study's target causal effect of race estimand has been offered as an explanation for why it failed to replicate in a study that used a different set of names. David J. Deming, Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz, "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study," *American Economic Review* 106, no. 3 (2016): 778–806. For more on the hypothesis that class might be confounding the original Bertrand and Mullainathan study results, see "Greg vs. Jamal: Why Didn't Bertrand and Mullainathan (2004) Replicate?", Uris Simonsohn in *Data Colada: Thinking about evidence, and vice versa*, http://datacolada.org/51. I thank Liam Bright for suggesting I include this debate.

I will return to this objection more thoroughly later on in this chapter. But for now I will note that this response foreshadows a crucial question about what it is for a race to act as a cause. Whether something is a *confounder* of the causal effect of race or a *part* of the causal effect of race will depend on one's account of what it is for race to act as a cause. That is, the question of whether signaling *low socioeconomic status* "confounds" the causal effect of *being Black* on callback outcomes cannot be answered prior to figuring what it is for *race* to cause callbacks in the first place.

It bears noting that three other studies have successfully replicated the Bertrand and Mullainathan results. Amanda Agan and Sonja Starr, "The Effect of Criminal Records on Access to Employment," American Economic Review 107, no. 5 (2017): 560–564; John M. Nunley, Adam Pugh, Nicholas Romero, and R. Alan Seals, "Unemployment, Underemployment, and Employment Opportunities: Results from a Correspondence Audit of the Labor Market for College Graduates," *ILR Review* 70, no. 3 (2017): 642–669; Nicolas Jacquemet and Constantine Yannelis, "Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market," *Labour Economics* 19, no. 6 (2012): 824–832.

Bertrand and Mullainathan's experiment is a now classic example of a correspondence study, designed to assess the extent of differential treatment by race in the labor market.[51] Their finding that Greg fared much better than Jamal with employers—on average 50% better across all the résumés that were sent out—has broadly been taken to prove the existence of substantial racial discrimination in the labor market. The conclusion that being perceived as Black is a disadvantage when looking for work is no doubt true, but the explanation for how the study contributes to it is less straightforward than most interpreters have appreciated. In particular, I want to press on the assumption that the correspondence study presents such forceful evidence in its favor *because of its approximation of the idealized controlled experiment*. For it effaces the subtle methodological choices that social scientists make to avoid study designs that, while still similar in setup to the letter of controlled experiments, would resist rather than support causal interpretation.

I want to suggest that despite the intuitive appeal of the experiment's design, there is a puzzle in figuring how studies of this kind are supposed to work to deliver causal verdicts about salient social categories. To get to it, I present two cases, audit studies of my own invention, which look to investigate whether being taken to be *male* rather than *female* is an advantage in job interviews. Note that the inquiry this time will require more than just constructing matched résumés of fictitious applicants. My job seekers will be real live "auditors" of firms' interview processes and so will appear in in-person interviews. This setup requires that they be "matched" not only on their work histories and various employment-related credentials but also in everything they say and do over the course of their interviews.[52]

---

[51] Correspondence studies are conducted with fictitious applicants rather than live in-person auditors. Differences between correspondence studies and audit studies are important to track in methodological discussions of discrimination studies, but these matters are not so relevant for the purposes of my argument in this Chapter, and I will often elide the distinction between the studies.

[52] Although the particular job interview audit study designs that I put forth in this Chapter have not, as far as I know, been tested, large-scale audit studies with trained matched pair auditors probing discrimination in interview processes have been conducted. In 1989 and 1990, the Urban Institute ran two sets of employment audit studies probing discrimination in the hiring process against Hispanic job seekers and Black job seekers in Chicago and San Diego and in

Before I introduce my cases, I want to acknowledge from the outset several differences between an audit study that features actors portraying male and female interviewees and a correspondence study with Greg and Jamal résumés that matter to how social scientists interpret experimental results. Most notable among these is the quality of the "match" in the former study compared to that of the latter. Real-life auditors will invariably differ across many dimensions—their heights, tones of voice, whether they have dangling or attached earlobes, whether or not their personalities seem warm and inviting, and so on—some of which may matter to their interview outcomes. A common worry is that these unmatched characteristics across the group of male and female actors may make for differences that threaten the study's results as speaking to the interviewer's response to the (assumed) sex status of the interviewee. This in turn explains why it is commonly thought that a résumé correspondence study presents a better means of probing causation. For these studies can safely get around the problem by matching Greg's and Jamal's résumés letter-for-letter, thereby escaping the specter of causally confounding target inquiry into the causal effect of race.[53]

A concern about what it takes for a matching to be done "well" gets to the core of the puzzle that audit studies present, and so I will return to the matter shortly, but for now allow me two brief remarks. First, it bears noting that it is simply not true that a correspondence study design makes Greg and Jamal identical across every dimension but their assumed racial status. Greg Baker and Jamal Jones vary in where they sit in alphabetical order; their names are of different lengths;

Chicago and DC respectively. These studies investigated the effect of race on having one's application accepted, being invited for an interview, and being offered a job. Auditors who were invited for interviews "were trained to behave as similarly as possible in an interview setting" so to be matched as closely as was possible in their interview performances. Margery Austin Turner, Michael Fix, and Raymond J. Struyk, *Opportunities denied, opportunities diminished: Racial discrimination in hiring* (Washington, D.C.: The Urban Institute Press, 1991), 1.

[53] For discussion of the relative merits of correspondence and audit studies, see Section 2.2 of Marianne Bertrand and Esther Duflo, "Field Experiments on Discrimination," in *Handbook of economic field experiments Vol. 1*, ed. Abhijit Vinayak Banerjee and Esther Duflo (Amsterdam: North-Holland, 2017), 309-393 and Jonathan Guryan and Kerwin Kofi Charles, "Taste-based or statistical discrimination: The economics of discrimination returns to its roots," *The Economic Journal* 123, no. 572 (2013): F417–F432.

'Greg' is monosyllabic, whereas 'Jamal' is polysyllabic; and so on. Nevertheless, in designing and interpreting the résumé correspondence study, we take ourselves to be licensed to assume that those features that are not matched do not make for causally relevant differences that will affect the results of the causal inquiry into race. After all, so long as we cannot dissolve all differences across matched pair candidates, the experimenter who proceeds anyway with her causal inquiry must assume those differences that remain do not undercut her analysis. So the same must be assumed of my interview audit studies. For each, variation on those fronts we cannot control for—heights, tones of voice, earlobe status—let's stipulate, will be causally irrelevant to job interview outcomes. With respect to the causally relevant factors, the pair of contrast cases differ only in the sex of the interviewee. Second, in practice, audit and correspondence studies are conducted with a great number of matched real-life or fictitious candidates with many different characteristics which the researcher hopes are randomized so that the two pools are not systematically different in some causally significant respect. I speak about my cases as though there is only a single pair of candidates at issue in each study, i.e., as though we are concerned with only what happens to two particular matched candidates who are differently raced, Greg and Jamal. I make this simplification, because my aim is to get to the logical core of the experimental method, which rests on a comparison of two contrast cases; social scientific efforts to randomize and work with large sample sizes which are motivated by practical and epistemological concerns are not features of these studies that I take to be important to replicate in my discussion.

With all that, the logic that undergirds my cases should mirror that which supports "real" social scientific audit and correspondence studies like Bertrand and Mullainathan's. My hope is that once we've discussed these cases, we'll see that the lessons learned will also apply to Greg and Jamal as well. So without further ado:

AUDIT STUDY I: SKIRT INTERVIEW

> Two actors, one taken to be male and the other female, present identical résumés, answer interviewer questions identically, and affect the same tone, mannerisms, and general personality traits (as best as they can). The male actor also dons the same dress and wears the same facial makeup as the female actor; both actors wear skirts and facial makeup to their interviews.

As already noted, let's suppose that those observable characteristics of the auditors that are not matched by the experimenters are causally irrelevant to interviewer decision-making, such that all that differs that is causally significant is the perceived sex of the candidate. Then, it would seem that if one assents to the logical core of the idealized controlled experiment, one ought to take it that differences in interview outcome across the contrasts in AUDIT STUDY I are constitutive of the causal effect of sex on interviews.

Yet, despite the easy appeal of audit studies, I suspect that most would be ill-inclined to accept results from AUDIT STUDY I as revealing the causal role that perceived sex plays in interviewer judgments. And the reason seems to be this: The proposed setup duplicates the *dress* and *facial makeup* across the two interviewees, and this intuitively seems to be the wrong approach to designing the study. Duplicating these features across contrasts seems to *confound* rather than aid in isolating the effect of sex the experimenter is after.

The concern that making the male actor don the same dress and facial makeup serves to *undercut* the match seems to me well-placed. For in honoring *this* duplication, the study does not in fact make good on the goal of keeping "everything else" that is causally relevant the same across the two actors. One actor in a skirt portrays a gender conforming female job candidate; the other, a gender non-conforming male candidate. And this difference, we know, matters for how the interviewees will be received by the interviewer. So any difference observed in job interview outcomes could be due to *that* change—a change in the causal factor that is *gender conformity*, not the change in the causal factor of *sex* that we were after.

The setup in AUDIT STUDY I, in other words, fails to ensure a match between the change to some scene that is *targeted* by an experimenter who assigns a "perceived sex status" and the change(s) that are in fact rung in by that assignment. When both interviewees wear skirts and facial make-up, assigning different perceived sex statuses entails another difference across the two candidates: whether each displays gender-conforming behaviors. And this difference in how a skirt is received by an interviewer when worn by a man compared to how it is received when worn by a woman now risks confounding target inquiry into the hypothesis that being *male* is an advantage in the labor market.

Of course, this troublesome difference could have been eliminated had the male-presenting actor been made to wear, say, slacks, instead of the female candidate's skirt. The thought generalizes. For matched pairs in an audit study to be made truly "identical" in all causally relevant respects but sex, all differences rung in to potentially causally relevant extrinsic factors such as gender conformity that might crop up across differently sexed interviewees and strike the notice of an interviewer should be neutralized. If this is right, then perhaps what we need is a study designed along the following lines:

> **AUDIT STUDY II: CONFIDENT MEN AND MILD-MANNERED WOMEN**
>
> Two actors, one taken to be male and the other female, present identical résumés and answer interviewer questions identically. To avoid confounding by perception of gender non-conforming status that may be triggered by, for example, setting identical styles of dress and facial makeup across the auditors, the social scientists look to make sure that both actors display traits that they take to be gender-conforming. For example, though the male actor presents as a confident and assertive candidate, the social scientists have the female actor portray as mild-mannered and demure, as a part of the effort to maintain what they take to be gender-conforming affect.

Whereas the experimenters who run AUDIT STUDY I duplicate all the features that the female actor had, giving them to the male actor—display the same affect, wear the same clothes, and so on—the experimenters running AUDIT STUDY II are wise to the different causal role that, say, *confidence* could play in the course of an interview when displayed by a female candidate compared to

the causal role that it plays when displayed by a male one. Their approach makes adjustments across the interviewee pair to account for these differences so not to trigger any of the "confounding" effects that had cropped up in AUDIT STUDY I. (Recall from the preceding chapter, the strategy of the Controls interventionist who looks to repair interventionist counterfactuals within which confounding effects crop up.) With a suite of modifications in place to strike out any changes to causally significant factors that may have been brought in tow by the change to assumed sex, the auditors in AUDIT STUDY II are thought to now present the correct contrasts from which differences in interview outcomes show the causal effect of sex.

But here too I anticipate some objections. For the whole point of social scientific audit studies is to approximate conditions of the idealized controlled experiment—to match candidates as closely as possible across their causally relevant features to isolate the effect of interest. By contrast, the setup in AUDIT STUDY II alters so many interviewee characteristics in order to keep the pair "matched" that we lose grip on the thought that these two interviewees differ only in sex. Now, they differ in some of the core personality traits that interviewers are precisely looking to gauge in the interview process. How could this be the right setup?

AUDIT STUDY I and AUDIT STUDY II present two poles of a spectrum of audit study designs that may be taken to show the causal effect of sex on interview outcomes. At one end, the design makes changes to all candidate characteristics that could be inflected by the social meaning of sex and could thus play distinct causal roles in the two interview contrasts. This is AUDIT STUDY II, which compares individuals different in some of the core personality traits that employers mean to judge in the interview process, a "match" that hardly seems to isolate the effect that is due to *sex*. At the other end of the spectrum is a design that changes only what is taken to be the *sine qua non* of sex

category membership[54], AUDIT STUDY I, yielding a comparison between gender-conforming and gender-transgressionist interviewees, an additional change that might now seem to mislead our target causal inquiry. Neither design seems to present the right pair of contrasts analysis of which gives an account of what it is for the sex of a candidate to causally affect interview outcomes.

And yet, barring any objections to the difference-making idea about causation at the heart of controlled experiments, it would seem that there must be some way of matching auditors to give the causal effect of sex. There must be some point along the spectrum of audit study designs that yields dependencies constitutive of the causal effects we are targeting. The question is where that balancing point is. Which characteristics should be duplicated across differently sexed interviewees? Which duplications bring in tow *other* causally relevant differences that must be adjusted to strike out potential confounding? So long as we reject the extremes presented in my cases, we are compelled to undertake this line-drawing exercise.

That no serious social scientist would pursue either AUDIT STUDY I or AUDIT STUDY II to investigate causal effects of sex already shows one thing: in practice, audit and correspondence studies are conducted with a sensitivity to features of causation in the social world and features of the causal roles of sex and race in particular that are uncaptured by considerations of abstract causal relevance alone. My cases are molded to two seemingly sensible ways of cashing out the reasoning which underwrites the audit study method, and yet, each strikes us forcefully as wrongheaded. The task is to figure what makes each faulty and what their respective shortcomings reveal about the construction of those audit studies which, at least on their face, appear not to have the same flaws.

[54] I want to note here that there are further ways to design an audit study that pushes in this direction of the spectrum. For example, one can set up a study in which transgender persons interview for the position.

§3. A diagnosis of the audit study puzzle

What reasoning about causation might justify the setup of AUDIT STUDY I? One might take it that for differently sexed auditors to be "matched" in the sense exemplified by the idealized controlled experiment, they must be identical in respects of the *intrinsic character* of all causal variables, except for sex status of course, on scene. This explanation would rationalize a setup that isolates the causal effect of sex by setting up auditors that mirror each other in not only what they say and how they say it—the acoustic quality of their responses should be identical, the cadence of their delivery, the frequency and timing of pauses, and so on—but also what they wear and how they wear it—the skirt, the facial makeup, and so on.

However suitable this approach might be in causal investigations of the natural world, it is plainly misled in the case of the social world, and the skirt is the smoking gun. The skirt on the body of an interviewee taken to be male is not the same causally efficacious thing as the skirt on the interviewee taken to be female. Even while no change has been made to the skirt *itself* as an article of clothing, it nevertheless changes qua causal factor across the two contrast cases. Or, to put the same point another way, how a skirt functions causally-speaking in the context of a job interview depends in part on who wears it—for example, whether the person is someone who is otherwise taken to be male or female. That is, the skirt is also characterized by a set of extrinsic properties, and the troublemaker here is one in particular: its status as a gender-conforming or gender non-conforming article of clothing. This makes for a causally significant difference between the male and female interviewees, too—one which undercuts AUDIT STUDY I's claim to "keeping all else equal" across the auditors *in the causally relevant sense*. And *this* interpretation must be what proponents of idealized

controlled experiments, including interventionists, mean when they put forth their method as a rigorous operationalization of the difference-making idea about causation.[55]

The conclusion that when probing causation in the social world, it is not a sufficient condition to "keep all else equal" to simply duplicate intrinsically defined variables across contrast cases, is a familiar moral of the previous chapter. But while the problem emerges in rather stark form to lay bare the error in AUDIT STUDY I, notice that the diagnosis applies more broadly. It applies even to Bertrand and Mullainathan's prudently designed résumé correspondence study. To return briefly to that study, the same lesson from the preceding chapter shows why there too, variables duplicated to be intrinsically identical across contrasts may not be playing the same causal roles on Greg's résumé as on Jamal's. Bertrand and Mullainathan do not create résumés noting that both job applicants graduated from, say, *Howard University*, for having attended a historically Black college functions rather differently as a bit of educational background depending on whether you are otherwise perceived to be white or Black.

But of course, that is not all. Having attended Harvard University, having a 3.8 GPA, having grown up in New Orleans, being fluent in German—all these lines on a résumé, matched letter-for-letter, *intrinsically* as it were, they may well be, still may not play the same *causal role* on Greg's as they do on Jamal's résumé. Take for example the seemingly more race-neutral fact of having attended Harvard University. Jamal's credential is much more likely to trigger a double-take than Greg's, whether as a mark of his truly outstanding abilities (having overcome such tremendous hurdles as a Black student growing up in New Orleans) or as a credential that raises suspicion as to his true merit (Is Jamal the beneficiary of affirmative action admissions policies?). And so there too, the claim that

[55] Another way about the same point, which foregrounds the problem of variable selection, is that gender conformity is another causal factor in the system, but one which is *extrinsically* defined, and so is left unaccounted for in the intrinsic character approach to variable selection. Failure to take it into account leads the interventionist to think, mistakenly, that she has intervened only on the causal variable Sex, when in fact, she has also changed the value of the causal variable Gender Conforming Status.

the setup of the correspondence study compares job candidates who are identical in all causally relevant respects but for race may not be true as one might have initially thought. While the hypothesis that being Black is a disadvantage in the labor market certainly stands, the correspondence study cannot be said to contribute to it *because* it identifies a difference in callback outcomes attributable to nothing else but race. For that claim to stand, the study must hold everything fixed in the *causally relevant sense* across the raced comparators, and that condition, I have argued, may not in fact have been achieved.[56]

If AUDIT STUDY I fails because it does not eliminate possible confounders that crop up in the form of extrinsic causal factors which change alongside changes to sex, then perhaps the right audit study is one that matches the *causal profile* of interviewees' characteristics. If a skirt on a female job candidate triggers in the interviewer the thought, "This candidate is dressed professionally and 'normally'," then a variable that plays the same causal role on a male candidate is one that also triggers the same, "This candidate is dressed professionally and 'normally'," thought. So, in place of the *skirt*, we swap in for the male candidate *slacks*. But if the problem is that *any* trait that is perceived differently and so is different causally-speaking depending on whether it appears in conjunction with an individual assumed to be male or one assumed to be female can confound inquiry into the causal role of sex, then experimenters must do more than trade skirts for slacks. The facial makeup must be scrubbed clean as well, for it too introduces a spurious causal effect when worn by the male candidate. And so also goes the confident affect the male candidate displays, for when the female

[56] Hence, I vehemently disagree with Guryan and Charles' claim that "correspondence studies sidestep the thorny issue of conceptualising the causal effect of gender or race" because they "are able to vary multiple attributes on the resumes randomly and independently. This is an important distinction between audit and correspondence studies. Whereas audit studies match full human beings, correspondence studies are able to create fictitious applicants that have any combination of attributes the researcher desires. Randomisation of these attributes can be independent, so it is possible to estimate the marginal effect of each of them and to estimate interactions" in "Taste-based or statistical discrimination: The economics of discrimination returns to its roots," F424, F423.

candidate presents as such, she is perceived quite differently—not so much as *capable* but as *bossy* and *unlikeable*. And so on. Rejecting AUDIT STUDY I for *these* reasons drives the causal inquirer straight into the arms of AUDIT STUDY II.[57]

I suspect that an approach that makes adjustments to all the traits that could trigger different interviewer responses when borne by a candidate taken to be female versus one taken to be male may strike as one that changes *too many* features across the two candidates. I want to identify this diagnosis with two distinct charges. Already mentioned is the loss in the credibility of the claim that the study contrasts auditors who are the same but for sex. The intuitive pull of the audit study qua an approximation of the idealized controlled experiment recedes further and further as the pair of interviewees are constructed to differ in more and more respects. A setup showing divergent callback rates across wholly different Greg and Jamal résumés—Greg attended Middlebury, grew up in Bethesda, Maryland, and speaks fluent French; Jamal attended Howard, grew up in D.C., and speaks fluent Haitian Creole—does not quite have the same force when presented as a study that shows the causal effect of just race on hiring prospects.

But second and more importantly for my purposes, the proposal to adjust everything possibly inflected by gender seems to change too much for a distinctively *normative*, as in *ethical*, reason. To see why it *overcorrects*, in the sense that the strategy pursued *controls away too much*, recall the

[57] It bears noting at this point that in actual audit studies conducted to probe discrimination on the basis of sex in extended dialogical interactions, experimenters *do* deliberately vary attributes across male and female auditors to ensure they "project similar class characteristics." See Ian Ayres, "Further evidence of discrimination in new car negotiations and estimates of its cause," *Michigan Law Review* 94, no. 1 (1995): 109–147, 113.
However, experimenters' decisions about whether observable traits should be varied or straightforwardly duplicated across differently sexed auditors to ensure "uniformity" are not explicated. For example, in Ian Ayres' audit study of the effect of race and sex on new car negotiations, dress was varied according to gender: "men wore polo or button-down shirts, slacks, and loafers" and "women wore straight skirts, blouses, minimal make-up, and flats." But when it came to training auditors in mock negotiations, the actors were instructed in a manner that aimed to achieve "uniformity in cadence and inflection" as well as in "nonverbal cues." For example, all auditors were trained to "avoid eye contact and not cross their arms" and to "feel comfortable with periods of silence." These choices cannot be justified only by reference to what attributes are likely to be received differently by salespersons depending on whether the buyer is a man or woman. For presumably, just as dress is likely to trigger in salespersons' different responses depending on the perceived sex of the buyer, so too do body language, nonverbal cues, and comfort with silence. Ian Ayres, "Fair driving: Gender and race discrimination in retail car negotiations," *Harvard Law Review* 104, no. 4 (1991): 817–872, 825, 826.

purpose of making the various modifications to the auditors in AUDIT STUDY II. The study is set up this way so to silence those causally significant differences across the differently sexed interviewees that would otherwise make for spurious causal confounding of the causal relevance of *sex*. (Recall what went wrong in duplicating the skirt across the two auditors in AUDIT STUDY I.) To silence those factors which play distinct causal roles across persons taken to be female versus those taken to be male is to strike them out of the ledger that tracks causal effects of sex in interviews. That is, after all, the *whole point* of making the modifications in the first place: to prevent the causal influence of non-sex factors from being laundered into and thereby mislead the story of how *sex* acts as a cause.

The thought that AUDIT STUDY II overcorrects, or alternatively, that it is completely fine, is to weigh in on the adequacy of this move—whether the setup is the right or wrong way to go about accounting for how an interviewer's perception of a candidate's sex status affects their judgment of them. My claim is that *any* such response, be it endorsement or rejection, is one that takes place in normative space and that draws on substantive moral and political reasoning about ethical notions and social phenomena like sexism, misogyny, gender ideologies, and the like. Now I want to emphasize: the point is not merely a psychological one. That is, I don't mean to suggest only that our intuitions about what makes for a good audit study design and what constitutes good inquiry into the causal effects of sex are pulled by our normative outlooks. Rather, my argument runs deeper and concerns the nature of the philosophical project of coming to a good metaphysical account of the causal structure of the social world and how salient social categories such as sex and race fit into it. As I will go on to argue, so long as we are committed to causal theorizing about sex and race, we are necessarily bound to ethical theorizing as well. No analysis of what it is for sex and race to be causes can extricate itself from engaging substantive normative reasoning.

Let us return to the spectrum of possible audit studies I invoked in the previous section. On one end lies AUDIT STUDY I, which changes just what one takes to be the *sine qua non* of category membership; on the other is AUDIT STUDY II, which changes everything with meaning that is inflected by gender category. It bears repeating that neither comparison put forth isolates the causal effect of sex as sought by the idealized controlled experiment or as imagined by the picture of surgical intervention at the heart of interventionism. For in both cases, the change to sex status entails changes to other causal factors as well. In the case of AUDIT STUDY I, it is the gender conformity of the candidates; in the case of AUDIT STUDY II, it is the whole range of features that are known to be received differently by the interviewer depending on the assumed sex of the candidate and are thus deliberately altered by the experimenters. In light of that fact, what are the grounds for preferring any point along the spectrum of possible audit study designs to any other in an analysis of the causal effects of sex? My answer is that any such choice draws on normative thinking. Allow me first to start by presenting something of an impressionistic picture that will, I hope, give a sense of what I mean here, before going on to defend the claim more thoroughly.

Suppose you are a social scientist designing an audit study investigating the causal effect of sex in job interviews. You are surveying your options. Should your study be more like AUDIT STUDY I, or should it lean towards the setup proffered in AUDIT STUDY II? To answer that question, you must first consider what it is to slide along the spectrum of possible audit study designs. What exactly are the choices that you face? Well, as you start from AUDIT STUDY I and head towards AUDIT STUDY II, you are deciding what characteristics of your pair of differently sexed auditors should be matched in the sense of being held to be *intrinsically identical* and which should be "adjusted" so that the characteristic has the *same causal profile* or *plays the same causal role* in the interviewer's judgment process. The intuitive reaction to AUDIT STUDY I, recall, was that it was a mistake to set up the study so to maintain the intrinsic character of the skirt and facial makeup on

the candidate taken to be male, and that adjustments are called for to the experimental design to make sure that his attire and overall presentation do not play a substantially different causal role in his interview compared to the role that his counterpart's attire and overall presentation play in hers. The decision to put forth a pair of contrasts that "adjusts for" attire and presentation to isolate the causal effect of sex status is one that takes a step away from AUDIT STUDY I and towards AUDIT STUDY II.

That judgment, of course, does not settle it, and you march along the spectrum, contemplating whether a certain variable should be "controlled for" in the sense of being duplicated across the two auditors and left in its intrinsic state or "adjusted for" so that its changed causal profile across contrast casts does not *confound* the causal effect of interest. At some point, you come across some attribute which plays a different causal role in the interview when it appears in conjunction with the candidate taken to be sexed female compared to when it appears on the candidate taken to be male, and you face a choice: do the interviewer's divergent responses to *confident affect* and *parental status* constitute an *effect of sex status* on interview outcomes? As was the case with the features *skirt* and *facial makeup*, the matter at hand is not whether the interviewer's different responses are inflected by the candidate's assumed sex status. The answer to that question, just as it was in AUDIT STUDY I, is a full-throated "yes." But as we saw in the case of AUDIT STUDY I, not all sets of differences constitute the effect of sex; there, the thought was that such differences *confounded* rather than *constituted* the effect of sex on interview outcomes. And so here you must ask anew: do different responses to *these factors* confound or constitute effects of sex?

Suppose at this juncture, you have the following thought, "Hold on, I don't want to *adjust away* the fact that when the female interviewee responds to the interviewer's questions with the same tone of confidence, she is perceived as a bossy know-it-all rather than a capable candidate. Nor do I want to adjust for the fact that her status as a parent raises worries about her ability to prioritize her

obligations at work; whereas for her male counterpart, parental status is judged to be a sign of maturity and well-roundedness. The fact that she is perceived as a bossy know-it-all who may not prioritize her work because of her childcare responsibilities is a part of the causal operation of sex in the employment arena. To adjust for these facts is to drop them out of that story. These factors and their effects matter for getting how sex fits into the causal structure of the social world *right*."

Here, your causal analysis of sex in interviews takes from your substantive moral and political thinking on such key normative matters such as how misogyny works, what constitutes gender-based discrimination, and what the scope of gender justice is. Your reasons stem from your belief that the disproportionate burdens of care work and social reproduction matter not only for the gendered division of labor in the home but also—insofar as they systematically present impediments to women's entering, reentering, and staying in the workforce—to the goings-on of the labor market and the workplace. You know that so long as paid labor is the primary means of attaining one's livelihood, differential access to employment entrenches inequalities in men and women's material well-being, which in turn further constrain the set of choices available to women, makes women less able to enjoy the benefits of social cooperation and more vulnerable to exploitation and domination, and undercuts their social and political equality. The structural explanation you have in mind shows unremunerated care work and gendered caregiving norms to be a matter of gender justice—even if many choices made within that structure are "voluntary."[58] The connection between one's assumed sex status and being taken to be, often accurately so, disproportionately saddled by care duties in the home traces out, on your view, a way those sexed

[58] Feminist philosophers have argued that the gendered division of labor in the home has implications for women's experiences in the sphere of employment and in turn for their social and political equality that make it a project object of concern, even for liberal conceptions of justice. See e.g., Anca Gheaus, "Gender justice," *Journal of Ethics & Social Philosophy* 6, no. 1 (2011): 1–24; Gina Schouten, *Liberalism, Neutrality, and the Gendered Division of Labor* (Oxford: Oxford University Press, 2019); Sally Haslanger, "What is a (social) structural explanation?," *Philosophical Studies* 173, no. 1 (2016): 113–130.

female are systematically subordinated in society. And it is for these reasons—whether or not you explicitly articulate your reasoning in this manner—that your analysis includes these factors in an accounting of how sex status acts as a cause in the social world.

One might now ask whether you *should* be freely drawing on this kind of substantive ethical reasoning and political commitment in an analysis of what it is for sex to be cause. Consider an opponent social scientist or causal theorist who challenges your doing so. He protests that your account of sex causation is infected, illicitly, by your normative analyses. Opening the door to these considerations, he says, makes for sham causal theorizing.

But what recourse does your opponent have? What alternative way of theorizing causation can he propose that avoids the substantive normative theorizing of which he accuses you? I think there can be none. After all, the claim put forth is that an interviewer's sense that a confident female applicant is a "bossy know-it-all" and the fact that her status of being a parent raises concerns about her commitment and capacity are psychological mechanisms and responses that are *a part* of how assumed sex status causally influences interview outcomes, for they are characteristic marks of misogyny and gender oppression. Any rejection of this claim which responds that sex does not act causally via these mechanisms can only be backed by normative argument about why not—why, for example, these are not workings of misogyny and gender injustice.

To go about the point in a different way, any rebuttal that posits a particular set of mechanisms by which race and sex act as causes must be defended with argument for why it should be that triggering *those* particular sets of social responses is how race and sex act causally. To take, as an example, a view prominent in the economics literature on causal effects of race as well as in the legal scholarship on racial discrimination, one might ask what makes the hostile attitude or ill-will of *racial animus* special such that when theorizing what it is for race to act as a cause, triggering an animus response should be the *only* way that race works? In light of the many mechanisms by which

race could causally influence outcomes, what supports an analysis that looks only for an animus mechanism, and excludes, for example, the causal significance of racial ideologies? Or, from the other end of things, one might wonder why an employer's assessment that a woman's caregiving responsibilities make her less able to flexibly adapt to work demands and fit in with the company's work culture constitutes a response to her sex. After all, granting for the sake of argument that these duties *do* make her less able to work on short notice and make her less amenable to the work culture, isn't the employer's judgment *not*, in the end, about her sex status?

But if there exists a range of plausible analyses of race causing on the table—and the many disagreements across competing theories of racism, racial discrimination, and racial injustice suggest there is—one can only endorse an animus-only causal analysis over others by making a move within normative space, that is by giving normative reasons for why we *ought* to endorse the animus-only view over all others. If a part of women's disadvantage in the labor market can be traced back to their disproportionate caregiving responsibilities making it difficult to take on inflexible work, then one must justify, as in provide normative argumentation for, why this fact should have no place in an accounting of how sex is causally relevant to employment outcomes. Insofar as racism and gender-based discrimination are a part of the stories of how race and sex causally influence social outcomes, then normative and explanatory considerations inevitably interact in the construction of theories of causation in the social world. For how one answers the question of what it is for race or sex to act as causes depends invariably on one's views about things like *how racism works* and *what constitutes gender discrimination*, matters about which there is considerable disagreement, emerging out of differences in substantive moral and political analysis.[59]

[59] The view that accounts of social categories and phenomena are themselves normative is not new. Alberto G. Urquidez argues that adopting a particular theory of 'racism' requiring weighing in on substantive normative matters in "What accounts of 'racism' do," *The Journal of Value Inquiry* 52, no. 4 (2018): 437–455.

Notice that my argument does not require that one be committed to a particular audit study design along the spectrum as the unique "right" pair of comparators that reveals the causal effect of sex. The same claim stands regarding the choice to dismiss any audit study designs along the spectrum as wrongheaded. The point is that *any* attempt to distinguish any design from any other as providing a better or worse accounting of the causal effects of sex requires some basis that cannot be reduced to a set of non-normative facts about dependencies. How one interprets the generic idealized controlled experiment dictate to "keep all else the same" across contrasts—whether it results in a choice to *control* or *adjust* for certain variables—is laden with ethical reasoning.

§4. An attempt to accommodate the lesson

If social scientific practices indeed evince a sort of wisdom about social causation, then not only is it a mistake to try to disentangle our causal theorizing from normative theorizing when it comes to developing an account of how race and sex figure in the causal structure of the social world, but furthermore, any such attempts to extricate moral and political thinking from causal thinking are in vain. As such, the fact that social scientists run certain audit studies and reject others as a means of causal inquiry into race and sex reveals that in fact normative theorizing is plainly part and parcel of our causal theorizing. And though my argument has centered on a puzzle presented by audit studies, the point generalizes beyond the narrow decision-making settings that these studies probe. Inquiry into the causal effects of race and sex at a broader societal level—the causal effects of

Esa Díaz-León argues that substantive normative considerations figure in debates about what race and gender terms mean even in our descriptive projects in "Woman as a Politically Significant Term: A Solution to the Puzzle," *Hypatia* 31, no. 2 (2016): 245–258.

race and sex on, say, various health outcomes—depends too on one's orientation towards matters of racial and gender injustice.[60]

But perhaps the sort of argument that largely reports from the frontlines of empirical social scientific inquiry can only go so far. Even if philosophical analyses of causation stand to benefit from taking a cue from scientific best practices, one might remain skeptical of a stronger philosophical claim. Can we really conclude from observations about social scientific inquiry anything about the *metaphysics* of causation?

My answer is yes, we can, and that in fact, the argument up to this point already makes for a serious challenge to a dominant view in the metaphysics of causation. Most everyone writing in the philosophy of causation believes that the causal structure of the world, that object of our scientific inquiry, is an objective structure constituted by natural non-normative relations. Christopher Hitchcock speaks rather directly into the camera on this: "[T]he causal structure is an objective feature of the system under investigation. In particular, the causal structure does not depend upon our interests or values… [I]nterests and values may influence which systems we choose to investigate, but they do not affect the causal structures of those systems."[61] But if my preceding argument convinces, then failure to engage moral and political thinking makes for a faulty analysis of how race and sex figure in the causal structure. Drawing on the normative is essential—lest we get the causal structure *wrong*.

This point leads naturally to a second lesson that bears on the metaphysics of causation. Insofar as an analysis of causation cannot accommodate normative inputs, it will be unable to limn a

[60] Studies of the causal effects of race and sex abound in the social sciences and are not limited to audit studies of discriminatory decision-making. Quantitative causal studies are many and varied—econometric regression techniques, causal inference using observational data based on the Rubin potential outcomes model, analyses based on structural causal models—but insofar as they follow the core logic of the idealized controlled experiment as a comparison of what happens in two contrast cases different *only* in the cause under study, my claim that normativity must make its way into the causal methodology applies in equal measure.

[61] Christopher Hitchcock, "Three Concepts of Causation," *Philosophy Compass* 2, no. 3 (2007): 508–516, 510.

good causal structure of the social world. This presents a few challenges to the standard interventionist analysis of causation. First, if approximations of idealized controlled experiments in the social sciences do in fact draw on normative resources, then interventionists who wish to maintain a tight connection to scientific practice must find a place for the normative in their theorizing as well. Doing so, however, is no small task. How can a theory that cashes out a causal claim as a claim about what happens under a highly specified *interventionist* counterfactual be reconciled with an analysis that reveals there to be a *spectrum* of different ways of filling in the content of the counterfactual contrast that is constitutive of causation? Incorporating the requisite considerations seems to deform the theory's core notion of an intervention as a change made only to the cause of interest that leaves in place all else causally relevant to the system.

The interventionist is unflinching in the face of this challenge and has a response at the ready. To the contrary, she retorts, a spectrum of counterfactual contrasts constitutive of the causal effect of sex counts in the theory's favor, for each option is underwritten by a different conception of the variable 'Sex' that is the target of manipulation. The many options laid out along the spectrum thus reveal a *virtue* of interventionism: that it ensures good causal hygiene by identifying when causal claims are fuzzy and imprecise. Interventionist analysis remedies the problem because, as Woodward puts it, the theory "helps to make such claims more determinate, clear, and precise"—it does so by making it explicit that they are to be understood in terms of one particular hypothetical experiment (which we specify) rather than another such experiment."[62] The puzzle of audit studies emerges because the causal inquiry into sex in interviews fails to spell out what exactly the target variable of 'Sex' is in a way that corresponds to a manipulation or intervention. Speaking directly to the case of sex and gender, Woodward writes, "From an interventionist perspective, the basic problem… [is] that the notion of a manipulation of or intervention on 'being a woman' or 'gender is unclear' [sic]

[62] James Woodward, "Methodology, Ontology, and Interventionism," 3589.

there are a number of different things that might be meant by this claim… causal claims concerning the effects of gender can be disambiguated by associating them with different claims about the outcomes of the manipulation of different candidate cause variables."[63]

The thought goes that if there are many different accounts of the category of sex or race, interventionism endorses for each a distinct set of manipulations as yielding a causal analysis of sex or race for that particular account of the category. This in turn explains the range of potential hypothetical experiments for probing causation. The interventionist can thereby have her cake and eat it too: causation is a matter of what would happen under an interventionist counterfactual, where the target of manipulation, the category of sex, is a category about which there is substantial disagreement. Should sex in the social world refer only to the possession of certain sex organs? Should sex status count visible secondary sex characteristics? Should it also encompass gender stereotypes? If so, which ones? The interventionist concedes that these questions matter to our theorizing about sex qua cause. But they matter in deciding among competing analyses of the category of sex, a step of the analysis that may be distinguished from the causal accounting itself. The charges against interventionism fail to make contact, for not only can the theory be reconciled with the existence of a spectrum of audit study designs, the kinds of normative analysis that I take to be integral to causal analysis, when properly conceived, is in fact a feature of theorizing about the target variable, which takes place *prior* to causal theorizing. So interventionism meets my challenge, and then some.

If the array of options for defining causal effects of sex arises out of a wide range of competing analyses of what sex is, then each audit study pairing could be said to give *a* causal effect of sex or a causal effect of *some conception* of sex. But the suggestion that each possible way of setting up the audit study corresponds to a different account of the category of sex seems to be specious

[63] James Woodward, "Methodology, Ontology, and Interventionism," 3590.

from the jump. The spectrum of contrast cases, recall, results from the different choices one could make about which causal factors should be duplicated to retain their same intrinsic character and which should be "adjusted for" to ensure they play the same causal role across candidates taken to be differently sexed. It is implausible that every which way of making such choices is underwritten by some account of what sex *is*. What account of sex takes the social category to be constituted by (gendered) affect but not include, say, (gendered) division of labor? What conceptualization of sex would call for an intervention that target changes to an individual's presentation of their face and hair but makes adjustments to their clothing choices to avoid confounding? The spectrum of audit study designs shows that every combinatorial possibility of "controlling for" and "adjusting for" causal factors generates a different analysis of sex causation in job interview settings. But these choices are not simply downstream of a generic theory of sex that is prior to and independent of the particular causal inquiry at hand that instructs the interventionist on which features to manipulate and which to adjust. The spectrum of audit study designs draws from substantive theorizing about how sex or gender might figure in a particular context; they does not emerge, as the interventionist suggests, solely out of the multiplicity of generic accounts of sex or gender.

This brings us to what I take to be the crux of the matter: whether an account of what sex is and an account of the causal role that sex plays in the social world can be disentangled at all. The interventionist's defense not only takes there to be such a division but furthermore, takes there to be a priority ordering to our analyses. There are, first, different accounts of what sex is or what race is. Then, resulting from this pluralism about what these social kinds are, flows various different counterfactual contrasts that are constitutive of causal effects of sex or race.

I want to suggest that this framing has things front to back. Probing the kinds of social patterns that accompany sex and racial status *can* help us come to better accounts of sex and race as social categories. If a good account of sex and race is one that is explanatorily powerful, then what

makes for a good social metaphysics will depend in part on the kinds of social phenomena we are looking to explain and what constitutes good explanations of those phenomena. And if these matters are themselves shaped by normative considerations, as I will in due course argue, then analyses of both social causation and of social kinds will depend on normative considerations, too.

Before moving on to the normative and explanatory pitfalls of accounts of race and sex causation that neglect the lessons drawn here from audit studies, allow me to make a brief remark on the relationship between our metaphysical theories of social causation and of social kinds. In this vicinity is a concern that constructing an account of a social kind as a functional kind that can explain certain social regularities amounts to a sort of circular theorizing. For if the social patterns that a good social account of sex will be able to illuminate are ones already taken to be in some sense "caused" by sex, then there will be a tight connection between our accounts of sex as a social kind and of sex causation in the social world. While I cannot settle here the precise nature of this relationship, one thing seems clear: filling in the best metaphysical account of gender, whatever it may be, does not suffice to resolve all questions about its causal role in various social systems. Or to put the matter another way, dispute about which candidate account of the category is best does not exhaust dispute about the causal effects it has in particular domains. What is at issue in theorizing the latter is not merely whether we should take the category to be a biological or social kind and of what sort; rather any analysis of what it is for race and sex to be a cause of some outcome must be sensitive to the particular causal system and explanandum at hand. When we ask what the causal effects of being taken to be female rather than male are in job interview processes, we are concerned with how the *interviewer responds* differentially to candidates that he takes to be differently sexed. Our attention is on the interviewer's divergent reactions to the candidates, which are classified as either being a part of the causal effect of sex or not. Normative considerations figure in this exercise of sorting which differences are due to sex and which are not, but they enter not via making revisions

to one's general metaphysical account of what sex is. Instead, normativity enters directly into the task of figuring what it is for sex to causally affect interviews. What it is for sex to act as a cause in interviews will depend in part on what is an instance of sexism and misogyny, and to decide on *that*, is just to weigh in substantively on a normative matter. And so normative thinking a part of the causal inquiry itself.

Hence, even while it may well be that moral and political considerations feature in both our metaphysical theories of social causation and of social kinds—and I contend that this is indeed the case—what those considerations are as well as how they figure in metaphysical theorizing are not identical in the two cases. If this is all right, there are distinctive aspects of theorizing causation of these social categories that are not reducible to the accounts of categories themselves. And so normative factors must find their place in causal analysis itself. Interventionism along with any other theories of causation that cannot accommodate these features make for deficient analyses of the causal structure of the social world—deficiencies to which I will now turn and discuss.

§5. The trouble with "intervening" on sex and race

I have argued that decisions about which pair of differently sexed counterparts, or which audit study design, sets up the proper contrast to elicit causal effects of sex draw on ethical considerations. But is that as far the lesson from the audit study puzzle can take us? That is, is the charge only that the standard idealized controlled experiment model of causation makes no room for normative inputs? Upon correcting this oversight by admitting the requisite normative considerations, can targeted manipulations in the vein of an interventionist theory generate a pair of contrasts that yield dependencies constitutive of causal effects, all the same?

First is the question of whether an amended analysis of causation to include moral and political thinking of this form may still be properly identified as an *interventionist* account—especially

if, as I have suggested, decisions to leave intrinsically be or adjust for causally relevant differences do not reduce to differences in the conceptualization of the category of sex or race itself. If not, then it bears asking what metaphysics of social causation can more readily absorb into its analysis the kinds of ethical considerations for which I have argued.

Setting this matter aside for now, I want to return to the general challenge that faces interventionist analysis when causal factors are extrinsic in nature. Those problems, discussed at a more general level in the previous chapter, show themselves in the specific case of the audit study probing causal effects of sex in interviews. Recall the Controls interventionist's proposal to repair counterfactuals that are beset by causal confounders introduced by extrinsic causal factors.[64] Her strategy is to control for the extra causally significant differences rung in alongside the target cause by making multiple interventions to the scene that will cancel out their spurious effects. The approach failed to resolve what goes wrong in my previous set of cases, I argued, because non-causal connections between variables make it impossible to cleanly strike out spurious causal structure without introducing further changes to it.

This dialectic—the Controls proposal and my rejoinder to it—can also be illustrated in the interview audit study. Consider another variant of it. Suppose Daniel and Eunice apply to a position, presenting identical résumés, answering interviewer questions identically, and affecting the same tone, mannerisms, and general personality traits. Over the course of their interviews, each reveals that they are expecting a child. This piece of information matters to the employer for two reasons. First, it signals to the employer the extent to which she can expect Daniel and Eunice to be able to throw themselves headlong into their work obligations and commit to long days on short notice. Second, it matters because of the company's parental leave policies. Each employee can take three months of leave within a year of the birth of their child. So upon hearing that both candidates are

[64] See Chapter 1 §3.2.

expecting children, the interviewer anticipates that whoever the new hire is, regardless of whether it is Daniel or Eunice, they will likely take three months of leave.

Suppose I think that employer judgments about a worker's capacity to meet and go the extra mile on their work obligations should not count in an analysis of the causal relevance of sex on hiring. For this criterion seems to be a perfectly reasonable, perhaps even the most important, basis for distinguishing job candidates.[65] Then, to guard against differences in the interviewer's judgment of candidates' work capacity "confounding" the target causal inquiry into sex, the thing to do is to make adjustments that will equalize Daniel and Eunice along the dimension of "work capacity" so that any difference in interview outcome reflect only responses to *sex* and not responses to *judgment of work capacity*. Suppose that the interviewer would judge Daniel and Eunice to have equal work capacity if Daniel were, just the same, expecting a child and if Eunice were *not* expecting a child. The interviewer judges non-child-expecting Eunice to be just as capable of carrying out the position's heavy responsibilities and unpredictable hours as can child-expecting Daniel.

Does this change now make Daniel and Eunice the right pair of contrasts? Certainly not. For the proposed change does not simply serve to cancel out a difference in the interviewer's judgment of work capacity; it also introduces a new difference between the two interviewees, one which makes a *causal* difference and thus risks misleading the target inquiry into the causal role of sex anew. Now the employer expects that Daniel will take three months of parental leave some time in the first year

[65] It does not matter for my purposes whether the employer's judgment that child-expecting Eunice will be less able than child-expecting Daniel to meet the position's heavy work responsibilities is rational or not—though certainly one might be more inclined to take the view that differential judgments about productivity should not figure in an analysis of the causal relevance of sex to a hiring decision if such judgments are rational or correct. Notably, economists James Heckman and Peter Siegelman have argued precisely to this effect: that perceived differences in productivity levels which slip through efforts at matching make for a serious weakness of audit studies, since employers may be detecting productive characteristics in candidates that are unaccounted for by the experimental design. Heckman and Siegelman, I gather, would not want to count an employer's correct judgment that a female employee's bearing disproportionate childcare responsibilities will affect her productivity on the job as an effect of sex. This anticipates remarks I will make in the following section about how audit studies substantiate claims of discrimination. James Heckman, "Detecting discrimination," *Journal of Economic Perspectives* 12, no. 2 (1998): 101–116; James J. Heckman and Peter Siegelman, "The Urban Institute Audit Studies: Their Methods and Findings," in *Clear and Convincing Evidence: Measurement of Discrimination in America*, eds. Michael Fix and Raymond J. Struyk (Washington, D.C.: The Urban Institute Press, 1993): 187–258.

of working, whereas Eunice will likely not. And *this* difference, we anticipate, will figure in the employer's hiring decision. We could, of course, try to eliminate it by furthering altering the profiles of Daniel and Eunice. Perhaps Daniel announces at the interview that he will not be taking his parental leave. Or perhaps Eunice reveals that she will need to take a leave for three months some time in her first year for some reason unrelated to childcare. But these options make for additional changes that causally matter too, leaving the Controls interventionist with work left to do, more differences to strike out with further changes.

The point is not whether there are other ways to precisely cancel out differences in judgments of Daniel's and Eunice's work capacity and whether the experimenters are creative enough to be able to set up a clever design that can do so. Insofar as the strategy undertaken follows the Controls proposal, the trouble for interventionism is general. Whenever a factor is causally relevant to some outcome in multiple ways, making adjustments to that factor with the aim of targeting just one of its causal pathways invariably entails changes to multiple of its causal roles. In this case, there is a non-causal connection between the interviewer's perception of child-expecting Eunice as a woman and her judgment of Eunice's work capacity, which makes it the case that a change to the former such that the interviewer now perceives, say, a man Daniel, entails changes to the latter causal factor. The case illustrates that *this* change, a change to *judgment of work capacity*, cannot be surgically removed while preserving all other causal factors, including *anticipated parental leave*, as interventionism requires. Nor, however, do I mean to suggest that we simply cannot distinguish the causal roles played by these distinct factors. I believe we can, but the question for the interventionist is: what are the interventionist counterfactuals that do the trick? What precisely is the content of the pair of counterfactuals that we should look to and evaluate to come to our causal conclusion?

Problems of this kind laid out by the Controls interventionist strategy crop up whenever manipulations look to selectively vary statuses such as sex or race in a causal system in the social world while preserving the causal role of factors that have the effects they do in virtue of their connection to sex or race. For an interventionist, such efforts at disentangling causal effects are the name of the game and yet, are in many cases, I have argued, impossible to realize.

Still, this is a mere surface blemish of a deeper problem with an account that defines sex causation by separating the causal roles of, say, *skirt-wearing* and *judgments of work capacity*, from that of *sex*. For even if one takes it that causally relevant differences in *skirt-wearing* and *judgments of work capacity* ought not be laundered into inquiry into the causal role of *sex* in affecting interview outcomes, interventionist analysis fails to note why such non-causal connections might yet matter to metaphysical inquiry into what it is for sex to act as a cause in the social world. And indeed, I want to suggest that they *do* matter, for both explanatory and normative reasons.

There are notable correlations in our world between a candidate's sex status and whether they wear skirts or slacks and whether they are able to work extended hours on short notice or not. How interviewers receive job candidates is shaped by these regularities: the fact that male candidates (tend to) wear slacks not skirts to interviews, the fact that those sexed female expecting children (tend to) have less flexibility to adapt to fast-changing schedules. So, in the actual social world, the causal role that sex plays in interviews is intimately tied up with things like dress and childcare norms. When an intervention is made to break from those regularities, the situation is abnormal both in the statistical sense and in the sense of deviating from some social expectation or standard, thus pulling the comparison and observed effect away from the causal regularities that sex typically figures in in the actual world. A causal account of how sex fits into the causal structure in affecting interview outcomes that is based on such regularity-breaking interventions is one that may not be

representative of how sex causally works in affecting interview outcomes more broadly and in other similar contexts.

This shortcoming is no small matter, for if one of the aims of constructing an account of the causal operation of sex is to come to an analysis that can *generalize* across situations and thereby make sense of other regularities that obtain between sex and other outcomes, then a metaphysics of causation that proceeds by treating as *distinct* sex and factors that stand in systematic regularity with sex status, will exhibit serious explanatory deficiencies. That much of social life is patterned along sex status is central to its causal story. An analysis that separates sex from its correlates will fail to make sense of this pattern and in turn will fail to generalize and illuminate other social phenomena wherein sex figures. It will simply not make for good explanatory social theory.

What is more, given that one of the key background interests we have in theorizing about the causal roles of categories like race and sex is to generate a well of insights from which our justice projects can draw, explanatory deficiencies of interventionism lead also to practical pitfalls. An analysis that fails to see the significance of the relationship *between* the causal role that sex plays and the causal roles that dress, affect, and judgments of capacity play in interviews will also fail to attend to normatively significant features of how sex figures in interview processes. For even if an analysis allows one to predict, accurately so, that female candidates who reveal their childcare responsibilities will be penalized for seeming to have less capacity to meet long and unpredictable hours at work, the theory cannot tell what is seriously incomplete about the claim that "the bad outcome depends on *revealing childcare responsibilities*." The same can be said of the male candidate who wears a skirts to a job interview and receives poor ratings. Even if the analysis suggests, let us grant accurately so, that intervening to alter the skirt to slacks will improve one's job prospects, interventionism does not have the resources to tell what makes this causal explanation deficient both explanatorily and normatively.

An account that treats the causal operation of sex as independent of the causal operation of skirts and of childcare responsibilities will fail to illuminate what about *gender*, rather than skirts and the existence of dependents as such, produces the outcome. And so it is unable to guide us in seeing what is socially unjust about the fact that some people wearing skirts are penalized in interview processes, whereas others are expected to wear skirts at their interviews, and what is morally troubling about an intervention that suggests that the former set of individuals simply not wear skirts if they want to get the job. After all, from the interventionist's perspective, the suggestion that a male-presenting job candidate not wear a skirt to the interview looks just like a suggestion to not show up to the interview wearing, say, a loud Hawaiian print shirt or a pair of flashy neon green pants. Of course, since skirts are an article of clothing embedded in our system of gender, a suggestion that a male-presenting candidate not wear a skirt is a suggestion that the candidate conform to an (assumed) sex status and so one that reinforces a system of gender conformity, undercutting all gender identities outside of the binary. This marks a significant normative difference between a suggestion against a male candidate wearing a skirt, compared to a suggestion that he not wear a loud Hawaiian print shirt. An analysis that suggests intervening to make "skirt-wearing" "slacks-wearing" fails to see disadvantages accrued because of skirt-wearing to be a matter of gender justice, and so it fails to guide towards other ways of intervening that will contribute to destabilizing rather than reinforcing reigning gender ideologies.

It bears reminding at this point what good concepts and categories in metaphysics should do for us. When we endeavor to construct a good metaphysical theory of social causation, we are looking to limn a causal structure of the social world that captures something of explanatory significance, something telling of the *systematicity* with which categories like race and sex figure in social reality—something, in Ted Sider's words, that is capable of "playing [an] explanatory role in

social theory."[66] I've argued that a causal structure articulated in interventionist terms is ill-equipped to achieve that basic explanatory goal with respect to race and sex, and so the interventionist's conception of race causation and sex causation fails on the grounds of good metaphysics.[67] The notion of an intervention that stands at the core of interventionism defines the causal relevance of race and sex by *separating* it from other causal variables in a system. But as we've seen, features like "Howard University" and "skirt" often have the causal effects they do because of the relation each stands to race and sex respectively. To disentangle the causal operation of these variables from race and sex is to misunderstand what race and sex as meaning-making social statuses are and to in turn, put forth an account of sex causation and race causation that misses out on important parts of the causal explanatory structure of the social world, which aside from being important to track for theoretical inquiry, matter too for our practical, moral, and political aims.[68]

Consider a final line of defense put forth by the interventionist, who counters that her preferred causal analysis of sex casts a wide net so to include the causal roles of various social factors that comprise gender. The claim is that *her* interventionist analysis of sex in the social world,

---

[66] Ted Sider, "Substantivity in Feminist Metaphysics," *Philosophical Studies* 174, no.1 (2017): 2467–78, 2473.

[67] I could have alternatively formulated metaphysics as concerned with "joint-carving," but given a tight connection between an entity's "joint-carvingness" and its explanatory power, the point is the same. See e.g., Ted Sider, *Writing the Book of the World*, (Oxford: Oxford University Press, 2011).

[68] If concepts earn their keep by making good on the aims of our explanatory inquiry, then what makes for good metaphysics will depend in part on what interests we have in pursuing a particular line of inquiry—interests which may depend also on our practical, moral, and political aims. Consider, as an analogy, the relationship between the field of human physiology and that of medicine and health. One aim of ours in studying human physiology is to satisfy our curiosity about the human body and how it works. Another aim is to be able to improve human health. An approach to studying human physiology that is successful at providing insights into the goings-on of the human body but is utterly ill-suited to the project of improving human health would be rather inadequate practically-speaking. And I gather it is uncontroversial to suggest that it would be a rather significant knock against the theory if it did not provide guidance in improving health—and especially if it suggested approaches to medicine that undermined health. I owe this point to Ned Hall.

Esa Díaz-León makes the same point in "Substantive metaphysical debates about gender and race: Verbal disputes and metaphysical deflationism," *Journal of Social Philosophy* 00 (2021): 1–19.

Sally Haslanger's metaphysics is centrally concerned with the notion of "aptness" of philosophical concepts—the extent to which they fit our various purposes of inquiry. See Sally Haslanger, *Resisting Reality: Social Construction and Social Critique,* (Oxford: Oxford University Press, 2012).

informed by her ethical commitments admittedly, is not vulnerable to the preceding charges of explanatory and normative deficiency.

My preceding discussion should make clear that I take no issue with an approach that draws on moral and political resources, and in fact, in my view, an adequate causal analysis of sex demands it. But I want to raise again the question of whether such an account of causation may be properly termed *interventionist*. I have suggested that sex status, as a marker of social difference, systematically correlates with many social factors and that a good account of how sex figures in the causal structure of the social world will be attentive to the regularities among these social factors and their causal roles. However, it seems to me that the closer a theory tends in this direction, the further it leans away from the interventionist's characteristic feature of making surgical changes to a target variable such that "everything else is held fixed." The more an account of sex causation draws on a theory of gender as a system constituted by a set of social interpretations and arrangements that affix to assumed sex status, the less it aligns with the core of interventionism which defines the causal relevance of sex by separating it from other causal variables in a system. Defining causation by surgical intervention is plainly at odds with the aims of a metaphysical inquiry into sex and race causation in the social world: both the explanatory project of illuminating how the social world works now as well as the normative project of theorizing a better one to come. And given the profound ways society is presently structured along these dimensions, these theoretical and practical aims are intertwined. Interventionism cannot do the job justice.

§6. Audit studies redux

If not by approximating the idealized controlled experiment, how do audit studies proffer strong evidence for causal claims about sex or race? Rather than serving to undercut their claim to

sound causal inquiry, clarifying how audit studies make *ethical* claims is in fact crucial to understanding how it is that audit studies can make causal ones.

I have argued in this chapter that no audit study design, no pair of differently raced or sexed counterparts no matter how carefully constructed, successfully isolates a difference that may be due "just to" race or sex as is the aim of the idealized controlled experiment. Instead, in putting forth a particular pair of differently raced or sexed counterparts, we draw on ethical reasoning, not considerations of causal relevance alone—to decide which features should be "controlled for" in the sense of being duplicated intrinsically or "adjusted for" in the sense of preserving the same causal role across counterparts.

But now arises a rather awkward question: having departed from the setup and logical core of the idealized controlled experiment, what *now* licenses reading off differences in some outcome in an audit study as telling of the causal relevance of race or sex? Having broken with the gold standard of causal methodology, on what basis can we still claim to be pursuing honest causal inquiry?

And yet, despite my entreaties to see auditor pairings as emerging from a particular normative position rather than a formal procedure that makes individuals identical but for race or sex, the intuitive pull of the audit study as a near-gold standard causal study still tempts. Knowing better then, should we try to resist their appeal? Perhaps surprisingly, my answer is that we should not, for as I will argue, audit study results *are* still causal studies of race and sex. Though the reason they present evidence of causation, and so forcefully at that, is because they put forth an *implicit normative argument* for when some outcome *wrongfully* varies with race or sex.

The point is made clearer when we return to how normativity figures in deciding whether a differential response to some feature should count as a *confounder* or a *component* of the causal effect of interest. Recall your thought process as a social scientist marching along the spectrum of possible audit study designs. At some point, you were deciding whether to adjust for the fact that the

confident female interviewee is taken to be a bossy know-it-all, and you had the thought that to do so would be to eliminate it from the causal story of how sex status affects interview outcomes. You realized that adjusting away this response was to strike it out of causal consideration for how those taken to be female are disadvantaged in hiring. And that doing so would put forth an incomplete picture of how sex figures in the causal structure of the social world. For, your thought was, the confident man and the confident woman *should* be received in the same way by the interviewer—presumably as *both* competent—and showing that the confident woman is not is important for it reveals a disadvantage of being taken to be female in interviews. It is precisely that fact—the fact that these two auditors *should* be treated the same and yet are not—that we are looking to show when we are probing the causal effect of sex in interviews.

What you have done is set forth a pair of comparators that, on your view, are substantively identical, *in spite of* a difference in sex status and all the differences entailed in the interviewer's perception of different sex status. And so these differently raced or sexed counterparts set forth a normative standard for *when* individuals *ought* to be treated equally. If this is right, then deviations from equal treatment show, per these standards, unfairness or discrimination *in the first instance*. That is, if audit study pairs propose a substantive moral standard of equality, then deviations are unfair *by reference* to that standard. The outcome is discriminatory *in virtue of* the fact that it fails to meet the proposed standard. Audit studies thereby furnish evidence for discrimination directly. But since racial and sex discrimination are a part of the story of how race and sex act as causes in the social world, audit study results do still probe the place of these social statuses in the causal structure but *indirectly* so: by first nominating a standard for what constitutes *fair treatment* across differently raced or sexed counterparts and then interpreting a failure to meet the standard as evidence of discrimination on the basis of race or sex—a case of *wrongful* causal influence of race and sex.

According to the interpretation on offer here, an audit study need not make any claim to causation in order to substantiate a claim of discrimination, because discrimination is not an essentially causal notion. It instead marks out cases when treatment systematically differs across racial or sex lines despite the fact that the candidates are proposed as being substantively equal in the sense of deserving normatively equal treatment. If an audit study's causal conclusions stand, it is not because it has managed to disentangle the causal effects of race or sex from the causal effects of non-race and non-sex factors. Rather, it is the fact that the study can directly show discrimination (non-causally) along with the implicit assumption that discrimination is one way that race and sex can act as causes that audit studies nevertheless probe causal effects. By providing evidence of racial or sex discrimination, they provide evidence of race or sex causation.

Standard interpretations of the relationship between audit study results and racial discrimination get this backwards. By and large, they suggest that audit studies proffer evidence of discrimination because discrimination *just is* a matter of causation, and so to the extent that audit studies show race or sex causation, they show race or sex discrimination.[69] This is why critics of audit studies as methods that provide solid evidence for discrimination challenge whether they do in fact successfully isolate causal effects.[70] Hence, in their seminal critical review of audit studies, economists James Heckman and Peter Siegelman write that the "main defect of matching methods" is that "they do not account for unobservables that affect outcomes."[71] The charge is that

---

[69] Recall Michael Gaddis' description of the relationship between audit studies and racial discrimination. Starting from a definition of racial discrimination as "differential treatment on the basis of race that disadvantages a racial group" (Blank et al. 2004: 39), Gaddis explain that audit studies provide evidence of discrimination by "manipulat[ing] the second part (race) to directly capture the first part (differential treatment) of the definition. Thus, by carefully controlling and counterbalancing all other variables in the experimental process, audit studies provide strong causal evidence of discrimination." Rebecca M. Blank, Marilyn Dabady, and Constance F. Citro, *Measuring Racial Discrimination* (Washington, DC: The National Academies Press, 2004); Gaddis, "An Introduction to Audit Studies in the Social Sciences," 9.

[70] For example, in a discussion of the weaknesses of audit studies, Bertrand and Duflo note that failing to perfectly match auditors such that the pair is "identical in all dimensions that might affect productivity in employers' eyes, except for the trait that is being manipulated" undercut their claims to having isolated just the trait of interest. The trouble is that even the most stringent matching efforts are "unlikely to erase the numerous differences that exist between the auditors in a pair." Bertrand and Duflo, "Field Experiments on Discrimination," 318.

[71] James J. Heckman and Peter Siegelman, "The Urban Institute Audit Studies: Their Methods and Findings," 188.

differences in causally relevant factors that the experimenter does not or cannot account for may mislead inquiry. The social scientist is a hunter of causes, and her job is to direct her aim as precisely as possible on her target—leave it to those on the normative side of the division of labor to decide which causal effects she should track down.

By contrast, if as I have argued a pair of contrasts can neither in theory nor in practice isolate the effect of *just* race or sex on some process of decision-making, then in proposing a standard by which to judge racial discrimination, the experimenter has already leapt the boundary to the side of the normative. Only his having done so is obscured by the widely shared moral intuitions of where he has landed. Greg and Jamal are owed equal rates of callback, *when they share identical résumé contents*. Daniel and Eunice ought to be evaluated equally well, *when they answer interview questions identically.* These audit studies put forth circumstances under which it is thought that Jamal *should not* suffer such a disadvantage in the labor market and Eunice *should not* be penalized in interview processes. The fact that many people agree that under these conditions, such disparities would be unfair explains why studies like Bertrand and Mullainathan's present such forceful evidence of discrimination. In a society characterized by stark racial inequalities in education access and unevenly distributed job opportunities exacerbated by racial segregation, that Jamals receive 50% fewer callbacks relative to Gregs *even when* applying for the same jobs with identical résumés shows the existence of substantial racial discrimination in the labor market, which in turn shows the causal significance of race in affecting employment outcomes.

My suggestion that the intuitive pull of the audit study derives from the rather common *moral* views of *when* Jamals *ought not* fare so much worse than Gregs also explains why critiques that aim to undercut the study on causal methodological grounds do not necessarily endanger its status as proving discrimination. Consider as an example of this line of criticism, Heckman and Siegelman's charge that a particular "Anglo/Hispanic audit study" suffered from problematic confounding

because "all the Hispanic testers in San Diego had facial hair and strong Hispanic accents."[72] Such an objection might yet leave us cold as a charge meant to challenge the study's claim to give evidence of discrimination. For even if employers were indeed responding negatively to the Hispanic auditors' Hispanic accents, I suspect that many would still take the study's results to show discrimination against Hispanic job applicants. What I want to suggest is that it is because we have this substantive moral view that we think that an audit study with these apparent imperfections are not invalidated on their basis. Still, they show discrimination, and for that reason, they show causation.[73]

The insight that an audit study's claims to causation go first through substantive ethical reasoning paves the way towards a radical revision to the way we go about causally theorizing about race and sex. If audit study pairs present a *substantive normative* standard of equality across differently raced or sexed individuals, rather than presenting individuals who are in fact identical in all respects but for race or sex, then the particular choices of matching reflect views about what makes two individuals substantively equal *despite* the fact that they are different in at least one significant respect: in their race or sex. Coming to an account of substantive equality given the social fact of racial or gendered difference, or an account of the conditions under which differently raced or sexed individuals ought to be treated similarly despite differences in race or sex, is a matter of ethical theorizing, which in turn requires social theorizing to figure what the categories of race and sex are as markers of social difference in our world. But once we see that it is on normative grounds that such standards of equality are put forth, it becomes clear that conditions approaching a perfect "match" need not be the only circumstances under which Jamals ought not be substantially

[72] Ibid., Ibid, 217.

[73] On this view, neither does rationality come into it. That is, the employer could be perfectly rational in making the judgments she does. Perhaps it is rational for her to judge the Hispanic candidate to have weaker soft skills (this was an "unobservable" causal factor for which the experimenter could not perfectly match). We can even suppose something further, that the employer is *right* in her judgments about soft skills. It really is the case that the Hispanic candidate has fewer job-relevant soft skills. Still it seems that this should not or at least *need* not undercut the audit study's findings of discrimination against Hispanics, and the judgment of whether or not it does is normative. This gives a straightforward sense in which economics definitions of discrimination as deviation from a "baseline" of rational behavior is normative.

disadvantaged relative to Gregs in job prospects. If the audit study instantiates a substantive normative standard of equality given racial difference, then there is no reason why the pair must be matched across every aspect of their résumé contents to tell of discrimination.

An example will be instructive. Consider a case in which Jamal has *less* work experience than Greg. One might take it that under these circumstances, Jamal *still* ought not face such substantially dimmer prospects than Greg when looking for employment. That is, one might take it that even under these conditions, a substantial disparity in callbacks would show discrimination. Of course, supporting this claim requires normative argument, but so too does Bertrand and Mullainathan's claim that Gregs and Jamals with identical résumés should receive equal callback outcomes—the only difference being that in that case, the normative grounds strike most as so obvious that it is fine to leave them implicit.

Still, it pays to make the reasoning explicit. Why, in the actual study, should Jamal not fare worse than Greg? Here goes one argument that could be made. Decisions about who should get a callback should be sensitive to applicant résumé contents, because résumés speak to candidates' qualifications. Since Greg and Jamal present identical résumés, they should be rated equally well with regards to their qualification for the job. And furthermore, nothing about the feature across which they are different—a difference in race—justifies *normatively-speaking* such a stark difference in callback outcomes given that they are equally qualified for the job. Hence, the observed differential shows discrimination.

Notice now that this is a normative argument through-and-through, and so nothing prevents us from making a similar point even when Jamals and Gregs *differ* in their résumés. Suppose Jamals have three months' less work experience than Gregs. One might still hold that decisions about who should get a callback should be sensitive to applicant résumé contents, because they speak to candidates' qualifications. But that in the case of Jamal and Greg who present nearly identical

résumés, the two features across which they are different—a difference in race and a difference in three months of work experience—do not justify such a stark difference in callback outcomes given that they present nearly identical résumés. Why? Again the claim might be that simply on *normative grounds*, such a modest difference in work history does not amount to a significant enough difference in the candidates' qualifications to justify such a disparity in callback rates. The disparity in this case too directly establishes discrimination.

In fact, nothing precludes arguing for a stronger claim still. Suppose that I concede that having three months less work experience should count against hiring someone. Jamal would indeed be a stronger candidate with three months more experience. Still, I think Jamals should not face such worse callback outcomes than Gregs, even when they have three months less work experience. If I am not to be inconsistent, what must my reasoning be? Well, if that difference in work experience is not to matter *when comparing Greg and Jamal*, then it must be that something about the other feature across which they are different—a difference in race—justifies treating them similarly, *despite* their different work experiences. The claim must be that given their difference in race, the three months work experience should not be held against Jamal *as a normative matter*, such that he faces such different prospects on the labor market. Now the grounds of this judgment are perhaps less obvious than the ones that backed Bertrand and Mullainathan's setup, but the point is that the task of meeting this burden is a matter of setting forth normative argumentation, not of finely executing causal methodology.[74]

[74] Many different kinds of normative grounds have been offered in the philosophical literature on affirmative action and preferential treatment in support of the judgment that Jamal should still not fare much worse than Greg despite his having less work experience. For example, in a society in which work opportunities are distributed unequally across racial lines because of pervasive racial discrimination, it is likely that Greg's more extensive employment experience does not indicate a better work ethic or even greater expertise in the area of work compared to Jamal. This might give reason to discount some of Greg's credentials. On arguments for preferential treatment, see e.g., Thomas Nagel, "Equal Treatment and Compensatory Discrimination," *Philosophy & Public Affairs* 2, no. 4 (1973): 348–363; James Rachels, "What People Deserve," in *Justice and Economic Distribution*, eds. John Arthur and William Shaw (Englewood Cliffs, New Jersey: Prentice-Hall, 1978), 150–163; Judith Jarvis Thomson, "Preferential Hiring," *Philosophy & Public Affairs* 2, no. 4 (1973): 364–384.

My suggestion, then, is that having resigned ourselves to failing the idealized controlled experiment, why not consciously break free of its bonds? So long as we can furnish normative arguments for the circumstances under which differential outcomes count as discrimination, we can design heretical "audit studies" with Gregs and Jamals instantiating those conditions and interpret their results in precisely the way we interpret orthodox audit studies: as showing evidence of discrimination as systematic deviations from equal treatment when Gregs and Jamals are similar or different in whatever respects such that they *ought* to be considered substantively identical.

The view according to which audit studies earn their keep not by isolating the causal effect of race but by directly setting forth a standard for when unequal outcomes constitute racial discrimination suggests further practical upshots in both the design of such studies and the interpretation of their results. Notably, whereas much debate about such studies has revolved around the quality of "match" that experimenters are able to achieve with their design, my proposal relocates the essential matter of properly studying discrimination from these particular methodological minutiae to getting clear on when differential outcomes across differently raced or sexed pairs are normatively unjustified in a way that constitutes discrimination on the basis of race or sex. Thus, the improved conditions of "control" in correspondence studies as compared with in-person audit studies do not necessarily make for better studies of discrimination. For while the currency of "control" matters on a causal interpretation of discrimination, it presents no necessary advantage on the wholly normative interpretation I have offered here.[75] There may in fact be reason

[75] Discussion of the relative virtues of correspondence studies compared to audit studies also reiterates that the task of isolating causal effects as the ultimate aim of such field experiments. See e.g., Marianne Bertrand and Esther Duflo's chapter on "Field Experiments on Discrimination":
"The correspondence method presents several advantages over the audit studies. First, because it relies on résumés or applications by fictitious people and not real people, one can be sure to generate strict comparability across groups for all information that is seen by the employers or landlords. This guarantees that any observed differences are *caused solely by the minority trait manipulation*." (emph. added). Bertrand and Duflo, "Field experiments on discrimination," 319.
For these reasons, correspondence studies have been taken to be a "significant methodological advance". Guryan and Kofi, "Taste-based or statistical discrimination: the economics of discrimination returns to its roots," F422.

to prefer audit studies to the extent that they may more naturally express the conditions under which Gregs and Jamals ought to be considered equal. For example, contra Heckman and Siegelman's critique that audit studies probing discrimination against Hispanics are lacking insofar as they fail to "unbundle accents, facial hair, and Hispanic status," audit studies may deliberately be set up so that auditors' features "bundle" together, perhaps in a way that reflects the distribution of those traits in the broader population.[76] Such a study would not only be unconcerned with the, in my view, contrived charge that effects of Hispanic accents "confound" inquiry into the effects of Hispanic status on job prospects, its results would likely reveal patterns of discrimination more telling of how being Hispanic causally influences outcomes in the social world more broadly. And on the basis of this greater explanatory power, we have good reason to take it to be a better exercise of causal inquiry at that—better, perhaps, than even the reputed gold standard.

§7. Trouble for interventionism as a methodological view

Philosophers have disputed what kind of a theory of causation that interventionism claims to offer. Some have interpreted and critiqued it on standard metaphysical grounds and so implicitly seem to be sizing it up as a metaphysical theory of causation. For commentators such as Michael Strevens and Eric Hiddleston, the theory's silences on the truth conditions or grounds for causal claims are glaring omissions.[77] Woodward in particular has emphatically denied such aspirations. Interventionist ideas are, instead, *methodological* proposals, guiding how we ought to make sense of and go about testing and making causal claims.[78] Since for him, the theory is essentially concerned with how best to achieve the aims of our causal inquiry, it is openly receptive to those contextual

[76] Heckman and Siegelman, "The Urban Institute Audit Studies," 226.
[77] Michael Strevens, "Review of *Making Things Happen," Philosophy and Phenomenological Research* 74, no. 1 (2007): 233–249; Eric Hiddleston, "Review of Woodward, *Making Things Happen*," *The Philosophical Review* 114, no. 4 (2005): 545–547.
[78] James Woodward, "Methodology, Ontology, and Interventionism."

and pragmatic factors which might make your typical metaphysician blush. Muddying the waters further still, some philosophers have despite Woodward's protestations defended interventionism's claim to being a legitimate metaphysical theory of causation.[79]

My argument in this chapter might be seen to take advantage of this ambiguity. I started with a puzzle about the setup of social scientific audit studies, showed the dilemma it poses for interventionism, then drew conclusions as to the theory's failure to provide an adequate metaphysical picture of causal structure. The synopsis, to recap, is this: in figuring what it is for social categories such as sex to act as causes, two natural ways of setting up counterfactual contrasts fail to show dependencies constitutive of causation. In the case of AUDIT STUDY I, the issue was that intervening to change only the *sine qua non* of category membership limns a causal structure with weak explanatory power. Separating out sex status from those traits that systematically correlate with it distorts the picture of the causal role that sex plays in the social world. The approach plainly makes a social ontological error about what the categories of sex and race are. The interventionist who moves toward AUDIT STUDY II adopts a thicker conception of what it is to be sexed in the social world by taking some correlations with sex to be constitutive of sex status itself. But what one hand gives, the other takes away. For though she might now get the social ontology right, she loses along the way her central notion of a localized surgical intervention, an unconfounded manipulation which changes *only* the cause of interest and leaves all else causally significant on scene undisturbed. In my resolution to the puzzle, I proposed that normative, as in ethical, theorizing is essential to causal theorizing about race and sex: to figure which factors should be changed or adjusted (and how) and which held to be intrinsically the same. The interventionist can find no place for these substantive moral and political considerations in her analysis at her peril.

[79] See e.g., Christopher Hitchcock, "Events and Times: A Case Study in Means-Ends Metaphysics," *Philosophical Studies* 160, no. 1 (2012): 79–96.

I levied the argument as a challenge to the adequacy of an account of the causal structure of the social world limned by interventionism. But does it have any weight for the interventionist happy to cleave from any metaphysical theses about causation and stick only to methodological proposals? To show that it does, I will draw out of the metaphysical critique lessons that implicate interventionism as a view on causal methodology.

Return, one final time, to the audit study. Audit studies are attractive not only for the intuitive appeal their results enjoy as causal results; they also seem to present an easy template for studying causation about social categories such as race and sex. Part of what is so compelling about the audit study puzzle is that it moves us to appreciate the serious stumbling blocks in trying to design such a study. Which features should be held fixed? Which adjusted? The obvious ways of doing it do not seem to work, but why exactly not? How then *do* we fill out that template? That these questions are quite tricky and their answers rather complicated shows that careful conceptualization of the manipulation at the heart of the interventionist's test is often taken for granted.

We need not stray far for an example. One of Woodward's own discussions of the different ways of conceiving of a manipulation of gender to interpret the causal claim "Being a woman causes one to be discriminated against in hiring" will do.[80] He gives the following as a candidate interpretation:

> (1) Intervening to change an employer's beliefs about the gender of an applicant will change that person's probability of being hired.[81]

"Note," Woodward writes, "that in the case of (1), the variable which is viewed as the target of the intervention (and the cause) is 'employer beliefs about gender' rather than gender itself,"[82]

[80] Woodward, "Methodology, Ontology, and Interventionism," 3590.
[81] Ibid., with renumbering.
[82] Ibid.

before going on to make the explicit connection to the audit study: "(1) is a claim that might be (and in fact has been tested) by, for example, submitting otherwise identical resumes in which only the gender of job applicants has been altered."[83]

Woodward, however, does not elaborate on what exactly an intervention on "beliefs about gender" amounts to. No doubt that a more careful specification of what *variable* the intervention targets gives a part of the story—and here Woodward claims a point for the interventionist framework: that it "forces one to be more precise about which variables are the intended causal relata"[84]—but failing to tell the *content* of these counterfactual contrasts, Woodward does not follow through on his own entreaty that we clarify our causal claims by specifying what exactly it is that we are our manipulating. Perhaps he thinks the answer is straightforward. We, of course, by now know much better. The audit study puzzle presses us to clarify what exactly an intervention on the target variable of "beliefs about gender" *consists in*. Without an answer, we are left in the dark about the conditions of the hypothetical experiment to which the causal claim supposedly corresponds. If interventionism is to provide fruitful methodological proposals, then it seems that at a minimum, the theory should be able to clear this bar.[85]

Equivocation on this front has contributed to two methodological pitfalls that plague causal studies of categories such as race and sex in practice. The first concerns the all-important question of which features in an exercise of causal inference should be controlled for and therefore "conditioned out" in order to identify some target causal estimand. In looking to identify the causal effect of race on police use of force, conditioning on a correlate with race, say a feature such as

[83] Ibid., 3591, with renumbering.
[84] Ibid., 3590.
[85] It seems to me likely that failure to notice that filling in the content of what it is to intervene on sex status or race is a non-trivial problem can in part be traced back to the cavalier attitude that interventionists have tended to take towards spelling out the truth conditions of their causal models. If this is right, then this would give another way in which metaphysical oversights can directly translate into methodological ones.

perceived level of suspicion, looks to dissociate judgments of suspicion from race "as such." The picture of what it is to manipulate race that rationalizes this statistical choice is one that takes it that intervening to change an individual's race (or if you prefer, a police officer's perception of an individual's race) does not entail a change to whether they are suspicious. So, the right pair of differently raced counterfactual contrasts will have the same "suspicion level."

I already gave my metaphysical reading of this conception of an intervention: it gets the social ontology of race wrong. For the more methodologically-inclined, the critique translates directly into a lesson in good social scientific causal analysis. To condition on the effect of perceiving an individual as suspicious when estimating the effect of race on policing outcomes, or to condition on the effect of having primary childcare obligations when estimating the effect of sex in employment, is to strike out precisely those differential social meanings and distributions of social benefits and burdens that constitute what races and genders are as social positions in a raced and gendered society. Conditioning them out vacates out those social factors that make these categories categories of social scientific concern in the first place. Once all these "other" correlates have been accounted for, what good is it to probe the causal roles of what remains?

In fact, we might ask, what *does* remain? Regarding the category concepts at the end of this distillation process, what remains of race is something akin to groupings of individuals based on similar sets of visible phenotypic features; what remains of sex is something like just possession of sex organs or observable secondary sex characteristics. But what can the causal effect of *these* categories, stripped of their social content, be? When, for example, the effect of police's perception of race on use of force is not permitted to include the judgments of suspicion or feelings of threat or so on brought in train, what is left? It seems that all that could possibly remain to be captured by causal studies of these thin social categories is a similarly thin affective negative response, something like plain animus as distaste. So our set of methodological proposals have carried us far afield from

the original categories of race and sex of interest and from the primary reasons for engaging in social scientific study of them in the first place. For it is *because* these categories are constituted by these social facts and so can do more than trigger reactions from distaste that we care to figure their causal significance at all.[86]

This leads to the second worry about interventionism as a methodological view, one I take to be more troubling, for it shows that the ideal of a localized surgical intervention may perhaps at present occupy too central of a role in social scientific causal inquiry. I have argued that if she wishes to get the social ontology right, the social scientist who studies the causal roles of race or sex must resist the urge to disentangle these social statues from their many correlates. If these features are a part of what it is to be sexed or raced, then the right thing to do methodologically is to not condition on these features. And yet to do so seems antithetical to the very core of the gold standard of causal inference practice as approximating conditions of the randomized controlled experiment. In such an experiment, the role of randomization is to make contrast groups on average the same along all other dimensions that might affect the outcome of interest. By ensuring that all significant differences in other variables across, say, men and women or Blacks and whites, are washed out—so no significant differences remain along, say, socioeconomic factors, and obligations inside the home, and so on—we can be sure that any difference in the outcome is due to the difference in the gender or race "treatment." The principle is the same in audit and correspondence studies: causal effects show themselves as differences in outcomes across Greg and Jamal *when each has graduated from Harvard*, across Daniel and Eunice *when each bears primary childcare responsibilities*.

So long as these remain the gold standard against which causal studies of race and sex are measured, I have scant hope that social scientific practice will embrace as proper causal inference

[86] I owe Issa Kohler-Hausmann for this point, which she articulates so forcefully in "Eddie Murphy and the Dangers of Counterfactual Causal Thinking," *Northwestern Law Review* 113, no. 5 (2019): 1163–1227.

methodology my proposal that we leave "bundled" correlates of the categories of race and sex. This is in significant part because social scientists seem to be powerfully guided by the picture of an exogenous manipulation made to change the state of just one variable, keeping all else equal. That is to say that on this point, Woodward is right: it appears that social scientists *do* identify causal claims with the outcomes of hypothetical experiments. Though this observation is, in my view, an unfortunate one. For the problem is that, when it comes to complexly constituted social categories such as race and sex, the hypothetical experiments imagined are just not ones right for investigating the causal dynamics of our social world. They are chronically prone to changing the topic, tend towards obscuring what social scientific causal studies intend to illuminate, and leave completely opaque the unique moral and political significance of their subject—liabilities which bear crucially on the practical, normative enterprise of methodology. If it is indeed in part the interventionist ideal which has led us astray here, then it seems to me foolish to look again towards interventionism to help guide us back.

Chapter 3. The Interventionist Causal Conception of Discrimination

§1. Introduction

In the early years of the Equal Employment Opportunity Commission (EEOC), the federal agency established under Title VII of the Civil Rights Act of 1964 to enforce prohibitions on employment discrimination, thousands of working women sent letters inquiring about their rights under the newly passed law. Women wrote in detailing their experiences with meager pay, harsh labor conditions, the unique burdens of shouldering both waged labor and unwaged domestic labor duties, in workplaces ranging from meatpacking plants to telephone call centers. Chronically understaffed and lacking an efficient means to process the diverse set of such claims, the EEOC began accumulating a backlog of cases from the day it opened its doors in July of 1965. By early 1967, the number of unresolved claims facing the agency had grown to 81,500.[87]

In response to mounting pressure from civil rights and women's activist groups who suspected the agency of willful neglect, the EEOC began an overhaul of their system of processing discrimination claims. The new approach required employers subject to Title VII to report hiring, promotional, compensation outcomes along with detailed job descriptions and demographic information. These data would then allow the agency to make comparisons across positions, companies, and industries and thereby detect systematic discriminatory practices affecting entire workforces. With the new statistics-based approach to investigation, the EEOC highlighted representation, or lack thereof, of women and racial minorities in certain positions as the prime indicator of employment discrimination. Major discrepancies in, say, the total number of women who applied for some job, and the total number of women who were hired for it, or stark disparities in the average pay for a position that was predominantly female and the pay for one that was predominantly male were *prima facie* grounds for discrimination. By the 1970s, the EEOC began

[87] Katherine Turk, *Equality on Trial* (Philadelphia: University of Pennsylvania Press, 2016), 31.

pursuing ambitious class-action lawsuits drawing mainly from statistical analyses, charging large corporate employers such as AT&T, General Motors, General Electric, and the nation's nine largest steel producers with discrimination under Title VII. One by one, the companies settled with the government without going to trial, instituting sweeping affirmative action hiring programs and agreeing to millions of dollars in back-pay settlements on a broad interpretation of what protection from employer discrimination requires.[88]

The last in the string of these cases, *EEOC v. Sears*, was the only one that would reach a court hearing and would become the case to show the legal prospects of substantiating a claim of sex discrimination by relying almost entirely on statistical analyses.[89] When the case first went to trial in 1986, EEOC's burden was to show with their statistics that sex status *caused* the disadvantages that women workers faced at Sears. Sears's burden was to rebut that causal claim. The job for the court in a case like this was to adjudicate which side was right on the matter of turning mere statistics into evidence of discrimination "because of" sex. Once the task was conceived of in this way—once showing sex discrimination became a matter of showing the causal effect of sex—proper causal inference methodology formed the core of the legal dispute. Litigation involving charges of a "pattern or practice" of discrimination has largely persisted in this vein since the mid-1980s.[90]

The broad acceptance of causal inference-based reasoning in *Sears* and dozens of discrimination cases since reflects the state of US disparate treatment law, which requires a showing of discriminatory intent on the part of the employer.[91] But the view that discrimination is a causal notion is one well-accepted outside of the evidentiary standards of the courtroom, by specialist and

---

[88] Turk, *Equality on Trial*, 39–40.
[89] Equal Employment Opportunity Commission v. Sears, Roebuck & Co., 628 F. Supp. 1264 (N.D. Ill. 1986); Equal Employment Opportunity Commission v. Sears, Roebuck & Co. 839 F.2d 302 (7th Cir. 1988).
[90] The Supreme Court put forth a statistics-based method for proving a "pattern or practice" of discrimination, or systemic disparate treatment, in International Brotherhood of Teamsters v. United States, 431 U.S. 324, 340, n.20 (1977).
[91] Though as with nearly all matters of legal interpretation, there is dispute among legal scholars about whether disparate treatment does require a showing of intentional discrimination or some bad mental state. Stephen M. Rich argues that it does not in "Against Prejudice," *George Washington Law Review* 80, no. 1 (2011): 45–46.

lay interpretations of the concept alike. To say that someone discriminates on the basis of sex or race is to say that sex or race makes a difference to what that person does. To say that someone was discriminated against because of their sex or race is to say that that person's sex or race made a difference to how they were treated. Legal and social scientific inquiry is preoccupied with these empirical features of discrimination. The central questions for scholars in these fields concern the ethical notion as a *phenomenon*. Something happens: was it an instance of discrimination? How should we go about detecting or testing for discrimination?

Philosophers of discrimination, by contrast, have tended to focus on the moralized concept, analysis of which is primarily concerned with the wrongmaking feature of discrimination.[92] Many such theorists offer reasons-based accounts, which scrutinize the deliberative process leading up to the discriminatory treatment. Discrimination on the basis of X happens when X wrongfully constitutes part of the grounds or reasons on which decisions are made or otherwise figures in the wrong way in the chain of mental states that leads up to the adverse outcome.

Reasons-based analysis may easily apply to the paradigmatic cases of discrimination involving explicit targeting of groups (e.g., "Whites Only" signs in shopfront windows) but quickly runs into trouble when figuring more complex cases. Cases of so-called "indirect," institutional, and algorithmic discrimination evade straightforward explanation by reference to the possession of intentions to act on the basis of certain traits or the existence of mental states that might be "biased." Unwilling to bite these bullets and write them out of the scope of the concept from the jump, some philosophers have argued that discrimination because of sex or race need not be accompanied by a chain of bad mental states. Sophia Moreau suggests that the "because of" clause

[92] David Brooks, "What is discrimination discrimination?," *Philosophical Papers* 11, no. 1 (1982): 15–30; Deborah Hellman, *When is Discrimination Wrong?* (Cambridge: Harvard University Press, 2008); Sophia Moreau, *Faces of Inequality: A Theory of Wrongful Discrimination* (Oxford: Oxford University Press, 2020); Benjamin Eidelson, *Discrimination as Disrespect* (Oxford: Oxford University Press, 2015).

in discrimination analyses could "alternatively refer to the causal chain that runs from the practice to the discriminatee via a particular trait of hers, where the policy would not have disadvantaged her had she not possessed that trait."[93] This causal analysis expands the concept of discrimination such that bad thoughts are not the wrong or harm at its heart and moreover, in a way that also unifies the seemingly disparate cases of discrimination.[94]

Rarely is the precise account of causation a point of emphasis within such analyses of discrimination, but something of an interventionist picture of what it is for sex or race to be a cause seems to hold sway in these accounts. On this analysis, the operative question is whether the (potential) discriminator would have taken the same adverse action had the discriminatee not had the "trait" in question but everything else about them remained the same.[95] This approach of distinguishing the trait's causal relevance to the outcome from that of all the other factors also scales up to analyses of discriminatory practices and systems. For example, discussions about the role of workplace discrimination in the enduring gender wage gap frequently break down into disputes about the causal influence of *sex* as opposed to the effects of other pay-related variables that are correlated with sex.[96]

This way of cashing out the causal analysis I take to be aligned with interventionism, as the counterfactual change that kicks off the altered history is a minimal change that alters only the trait but nothing else about the scene at hand.[97] Still there remains ambiguity as to which interventionist

---

[93] Moreau, *Faces of Inequality: A Theory of Wrongful Discrimination*, 20.
[94] Even better for the philosopher who takes reasons to be causes!
[95] In the law, this is termed "but-for" causation.
[96] See e.g., Mark J. Perry, "Details in BLS Report Suggest That the 'Gender Earnings Gap' Can Be Explained by Age, Marital Status, Children, Hours Worked, Etc.," (American Enterprise Institute, October 22, 2021), https://www.aei.org/carpe-diem/details-in-bls-report-suggest-that-most-of-the-gender-earnings-gap-is-explained-by-age-marital-status-children-hours-worked.
[97] Of course, upon *evaluating* the counterfactual to check whether the outcome of interest is different, many things will differ between what actually happens and what counterfactually happens. Moreover, although the analysis does not propose a particular manipulation that brings us to that counterfactual, I take it that it is nevertheless aligned with the interventionist's line of reasoning for the crucial matter is that the right contrast is one in which *only* the trait in question is altered. I am not, however, wed to the claim that the account of discrimination is an *interventionist* one. Any proponent of a counterfactual analysis of causation that picks out as the "right" counterfactual, a contrast that differs only in the

counterfactual gives the right contrast to test for discrimination. At what point in time is the intervention on the trait to take place? That is, when precisely are we supposed to imagine the intervention happening and the counterfactual clock running, as it were, to see whether the adverse outcome takes place? To test a claim of wage discrimination on the basis of sex, is the right contrast one that winds back to the moment after the employee's conception to change their sex, and then allows an altered history to unfold until the present moment to check for whether they are paid a different wage?[98] Alternatively, is it one in which the interventionist swoops in to alter the employer's beliefs about the employee's sex status at the point of their decision-making about compensation? Different still, is the right counterfactual contrast one in which a change is made to the social structure such that gender ceases to be a marker of social difference at all? This brings to light a further ambiguity about how to properly conceptualize the trait that is the target of intervention. Notice that each of the preceding counterfactuals refers to a distinct conceptualization of the variable "sex": the first takes sex to be a biological fact of the employee, the second refers to the employee's sex status as conceived of by the employer, the third refers to gender as a social structure.

The causal test for discrimination pursued in *Sears*, however, resolves these ambiguities and does pick out a particular counterfactual: one in which the imagined intervention on sex makes its change at the moment of or immediately preceding the employer's decision-making vis-à-vis the adverse action. Aligned with a conception of discrimination as an agential wrong, this counterfactual gauges whether the decision-maker's action is sensitive to the individual's possession of the trait in

target trait with "nothing else" causally relevant different is subject to my argument—self-identified interventionist or not.

[98] Woodward suggests this to be an option for interpreting what it is to say that sex causes some employment outcome in "Methodology, Ontology, and Interventionism," 3590. This causal interpretation of sex discrimination seems to me implausible on its face.

question (or, if you prefer, their belief about the individual's possession of the trait in question).[99] On what I am calling an *interventionist causal conception of discrimination*, the analysis looks to a counterfactual contrast in which an intervention is made to the potential discriminatee's trait (or, again, if you prefer, the potential discriminator's beliefs about the trait) at the moment right before the adverse action is decided. If they would have taken the adverse action all the same, then the trait was not causally relevant to their taking the action, and so there was no discrimination at play in the act. If their action would have been different, the trait was indeed causally relevant, and the discriminator indeed discriminated.

I argued in the preceding chapters that in the cases of sex and race, counterfactual contrasts in which hypothetical interventions replace an employer's beliefs about, say, sex status S with beliefs about different sex status S', are not well-defined.[100] My claim in Chapter 2 was that, contra the standard interventionist line, any analysis of what is for sex and race to be causally significant for some outcome pulls on moral and political considerations. The counterfactual contrast that is proffered to illuminate sex or race's causal role in some setup embeds within it substantive normative views about things like misogyny, racism, and more broadly, what is or is not wrong with how the social world is patterned along the lines of sex and race. If I am right that all such contrasts are shot through with ethical considerations, it follows that so too is the one put forth in the interventionist causal test for discrimination.

This chapter probes the *moral content* that is embedded in and enshrouded by the interventionist causal picture that undergirds the standard legal test for discrimination. Hence, I set aside the

[99] It therefore bears an affinity to the reasons-based analysis but is more expansive than it in its ability to capture unconscious or otherwise unintentional sensitivity to the trait. And so, while the problem of so-called "implicit bias" might get a free pass under a reasons-based view of discrimination, here it is caught.

[100] Or if it is, it is a matter that itself draws on normative considerations, including possibly what constitutes discrimination, making its application here as an input into analysis about what constitutes discrimination problematic because circular.

preceding chapters' focus on conceptual issues with interventionism applied to these social categories so to shift attention to the distinctively *normative* set of issues with adopting this particular causal conception of sex (and race) discrimination and correspondingly, sex (and race) equality.[101] Though it appears to simply take stock of the objective causal facts, the interventionist test, I argue, yields bad first-order moral and political consequences. The conception of sex assumed by the interventionist causal conception of discrimination is one that takes sex categories as marking a thin distinction without a social difference. Sex categories, on this view, do not mark out groups that are differently socially situated with respect to key life outcomes. Sex status does not bear on the distribution of key social benefits and burdens. The interventionist causal conception of discrimination based on this account of sex rules that proscription of sex discrimination in employment amounts to a prescription of sex *neutrality*, and hence, workplace equality between sexes is an equality based on *interchangeability* and *sameness* of sexes. This assumption of sameness across races and sexes, for one, flies in the face of social fact and is thus unjustified both empirically and at the level of the social ontology of these categories. But furthermore, as I will argue, its adoption is specious on normative grounds. The interventionist method for delivering verdicts on whether a given case or practice constitutes discrimination on the basis of some social category requires those who charge discrimination to meet an exceedingly onerous standard for proof to substantiate their claims and thereby systematically favors those accused of discrimination. The upshot is that the method yields deflationary conclusions about the causal significance of race and sex and in turn, race and sex discrimination.

[101] While philosophers are typically wary of equating *nondiscrimination* with *equality*, the concepts are much more closely tied in US law, as discrimination doctrine is enshrined in the Equal Protection Clause of the 14th amendment. In keeping with the legal discourse, I speak throughout this chapter of "equality" as it is commonly referred to in the legal sense.

While it is certainly a virtue of the interventionist analysis that it supplies a clear method for delivering verdicts about discrimination, its clear causal standard is also double-edged. Its test for adjudicating cases foregrounds questions of epistemic soundness of causal conclusions as the main point of dispute in conflict about discrimination—and not, notably, the ethical issues at stake in efforts to expand or delimit the scope of the concept of discrimination. The account thus exerts a discourse-shaping effect on substantive debate about the nature of discrimination. For example, in the battle over the extent to which prohibitions of sex discrimination in compensation requires employers to reevaluate wages set for feminized labor, interventionist causal reasoning about discrimination weighs in to short-circuit normative reasoning. The interventionist dictum that sex causal contrasts should be identical in all respects but sex all but determines the scope of protections against sex discrimination. If nondiscrimination on the basis of sex requires only that an intervention on an employee's sex status has no effect on how an employer sets their pay, then the law requires only that those sexed male and those sexed female in identical positions and who are identical in all other pay-relevant respects are remunerated equally. Differences in compensation across male-typed and female-typed jobs that are not identical or nearly identical are simply *ineligible* to be scrutinized for discrimination. The ethical matters at issue—debate, in this case, about the purpose and value of workplace sex equality, on one hand, and the purpose and value of preserving market-set valuations of labor, on the other—are preempted by the causal standard. That further normative debate is stalled by an analysis that favors a conception of sex (and race) equality as sex (and race) "blindness" shows interventionist causal reasoning about discrimination to have an *ideological effect*: it tends towards enacting material outcomes that uphold prevailing systems of injustice and furthermore does so in part by naturalizing those injustices as simply following from the objective causal facts.

The remainder of the paper proceeds as follows. I first provide an overview of this strand of interventionist reasoning about sex and race and show how it is applied to analyses of discrimination

on the basis of sex and race. My claim is that its operationalization as a legal test for discrimination generates two first-order normative upshots: first, the adoption of an equality-as-neutrality standard, a standard ill-suited to many of the social inequalities to which an adequate account of discrimination should be responsive, and second, an analysis that is systematically biased against findings of discrimination and deflationary of the causal significance of protected categories such as sex and race more broadly. This I argue in §2. In §3, I turn to the second-order effects of causal reasoning in this vein and argue that the core interventionist idea of causation occupies an agenda-setting role in discourse about discrimination by distracting from and stalling discussion of the substantive ethical matters at stake. My suggestion is that these normative upshots taken together make for a strong practical case against the interventionist causal account of discrimination. Finally, I close in §4 with some remarks on the virtues of viewing our ethical concepts as historical entities forged over time through the real goings-on of the real world. My suggestion is that this broadly materialist approach better illuminates what conceptual analysis about ethical concepts is and should be up to.

§2. First-order normative upshots of the interventionist conception

It will help to start by reviewing the core logic and background assumptions that underlie the interventionist picture of causation to bring out its normative character when operationalized as the legal test for discrimination. Recall that the notion of an intervention is meant to capture an idealized manipulation in an experiment. In the idealized controlled experiment, two contrasts are identical across "everything" but the factor under study, which is the target of intervention. A difference in some outcome across the two contrasts shows the causal relevance of the target factor, since everything else was "the same" at the moment of intervention. The surgical precision of the intervention ensures that *it* was the only difference at the time of intervention, and so only *it* among

the set of potential causes at that time could have been the difference-maker.[102] The target factor's causal status is thereby proven by a *method of elimination* which rules out the operation of all other causally relevant factors at the time of intervention at that might explain the effect in question. Statistics-based causal inference analysis follows the same negative strategy. As statistician turned leading causal inference methodologist Paul Holland puts it, "Before one leaps to a causal conclusion, one needs first to consider the other noncausal explanations and *eliminate them*."[103]

What must hold true for the method of elimination to draw the right causal conclusions? First there is an assumption about the nature of causal relevance. Consider an analysis of the causal relevance of factors X, Y, and Z at some time *t* to some outcome of interest at a later time *t'*. If the method of elimination has it that X takes what remains once Y's and Z's effects have been accounted for, then it must be assumed that Y's and Z's causal contributions to the outcome are distinct from X's. That explains why X does not get to claim any of their effects for its own. The causal relevance of X at *t* can be distinguished within and is to be parceled out of the combined effect of X, Y, and Z at *t*. Call this the *substantive conceptual* assumption. Second is a condition on the circumstances under which the method reliably works—given the preceding substantive picture of causation—to transform information about factors other than X into information about X: namely, only when one has struck out *all* other causal alternatives at time *t*. If X, Y, and Z are all three possibly causally relevant factors at *t* for the outcome of interest at *t'*, it is not enough to rule out just the causal operation of Y in order to pin an effect on X. For that leaves Z standing as perhaps the true cause of the effect. So, the method of elimination is sound only when one starts with the full set of causally relevant factors at *t* for the outcome of interest, since it is only then that one may be sure

[102] This procedure tracks John Stuart Mill's method of difference, first articulated in his *A System of Logic, Ratiocinative and Inductive. Vol. I*, Book III., Chapter VIII (London/New York: Longmans, Green, and Co., 1843).
[103] Paul W. Holland, "Causation and Race," *ETS Research Report Series* 2003, no. 1 (2003): 1.

to have zeroed in on the target causal factor. An interventionist pursuing her causal analysis in practice must assume to have satisfied this *epistemic* condition.

When plugged into analyses of sex and race discrimination, the background assumptions the causal interventionist about discrimination brings in tow here generate normatively significant upshots. In its preoccupation with the task of disentangling the causal relevance of sex and race from that of "other" potentially relevant factors at time *t* of the intervention, the substantive conceptual assumption imbues the causal test with moral and political content. In supporting deflationary conclusions about causal significance of sex and race in the social world, it severely narrows the scope of what constitutes sex and racial discrimination. Meanwhile uncertainty over whether the epistemic condition is satisfied presents a significant *practical* hurdle to substantiating claims of discrimination and is thereby biased against those who put forth charges of discrimination. What results is an interventionism-based legal standard, which though itself appearing to track only causal facts and thus be devoid of any normative commitments, is bent towards politically conservative verdicts of discrimination and accounts of equality. On the assumption that discrimination *would* be wrongful, the happy finding that there is little discrimination after all functions to exonerate the prevailing social system.[104]

§2.1 The Substantive Conceptual Assumption

Consider an interventionist's inquiry into the effect of sex on employer pay-setting. Her method is primarily concerned with distinguishing the causal relevance of the employee's sex to her pay from that of other causal factors at time *t* immediately prior to her employer's making their pay decision. Suppose one factor that figures in how an employer sets compensation is the type of work

[104] Alternatively, one might take there to be much injustice in the existing social system but take there to be few instances of the particular wrong that is *discrimination* having adopted a particularly narrow conception of it. Whether this option is attractive depends, I think, on practical considerations, which I discuss further in §4.

that the employee is engaged in. This factor is not equivalent to nor is it completely determined by the employee's sex. A pay-setting scheme that takes as input the kind of work that an employee is engaged in is clearly a different way about determining compensation from one that sets pay according to employee sex status. "Sex status" and "type of work" are therefore distinct inputs at time *t* into pay decision-making. Since interventionism looks to stave off the effects of potential causal confounders, standard interventionists would appear to favor an analysis of causal structure that distinguishes the two variables. After all, it is straightforward to conceive of a counterfactual state of affairs in which the value of one is altered while the other is fixed. Upon drawing this line between sex status and other "non-sex" causal factors, the interventionist takes the causal operation of the former variable to be distinct from that of the latter variables. Her analysis requires now that the causal significance of sex be disentangled from that of other causal factors. So, she will aim to control for, i.e., strike out the causal operation of, other factors at *t* that may influence determinations of pay, such as *type of work* and *schedule flexibility,* in her effort to drill down on the causal relevance of *sex*.

Interventionist analysis thus puts forth a pair of contrasts who are identical in all respects potentially causally relevant for employer pay-setting—the pair of individuals must perform the same type of work, require the same level of schedule flexibility from the employer, and so on—but who are differently sexed. Any difference in pay across the candidates gives the causal influence of sex, since, once again, sex was the only difference, so only it could have been the difference-maker. Statistical evidence proffered in the courtroom simply looks to operationalize the account. The causal inference expert scours the data for approximations of such sexed counterparts, compares their pay, taking differences as evidence of discrimination on the basis of sex.[105]

[105] Statistics-based causal inference usually proceeds by identifying causal estimands such as the Average Treatment Effect (ATE), which gives the effect of the category of interest averaged over some population of individuals who are taken to be suitable comparators. That this quantity is calculated by averaging over many individuals makes no

No doubt that explicitly sex-based pay-setting gives one means by which an employee could be a victim of sex discrimination, and the interventionist test vindicates this judgment. But an analysis that starts from a conceptualization of sex status that separates it from non-sex factors makes it difficult to countenance an account of sex discrimination that encompasses much more beyond this. Notably, it is no object to the interventionist, and so no object to causal experts in discrimination trials either, that there exist broad gender differences in, for example, the kind of work individuals are employed in and the distribution of burdens of social reproduction outside the arena of waged work. The protagonists of interventionist causal analysis are individuals who, *despite* differences in sex status, are otherwise "identical" at time *t*. The method tells of the causal significance of sex by comparing contrasts for which these differences across sex *groups* are expunged as much as possible. The interventionist's analysis in fact consciously turns from them and prevents them from being incorporated into her target causal findings. So in *Sears*, the defense claimed that pay differences between women and men were due to divergent "interests" for the higher-paid commission sales positions at the group level. The right comparisons to be made, they argued, are those between those women and men who showed equal interest in working nights, weekends, and irregular and overtime hours in order to determine the existence of sex discrimination in hiring and compensation for these positions that were more likely to come with such time burdens.[106] That is, notwithstanding the fact that women, given their position in the family, are on average less free to work nights, weekends, and take on irregular and overtime hours,

substantial difference to my points here. The need to sum over many individuals simply reflects epistemological constraints that make it impossible to calculate individual effect sizes in the social world. The core logic that underlies statistical methods that compute the ATE follows that of the controlled experiment and thus interventionism.

[106] *Sears*, 1308 survey results.

these social facts pertaining to the gender division of labor are to be set aside for the purposes of adjudicating sex discrimination.[107] The interventionism-based legal test for discrimination concurs.

Could an interventionist take on board these facts about gender as constituted by a set of social relations to reconceptualize the variables in her analysis so to count compensation based on, say "schedule flexibility," as an instance of compensation based on "sex." Certainly she could, but as I remarked in the previous chapter, the more one looks to expand the conception of sex that is the target of intervention, the further one seems to depart from the spirit of interventionism. For it is one of the special virtues of interventionism that the dictate to consider a surgical intervention that makes a minimal change to alter the variable of interest and nothing else picks out a well-defined counterfactual. Minimality seems to favor the hypothetical intervention that manipulates an employee's sex status without making changes to any other pay-relevant causal factors, which is after all not conceptually impossible, and which lowers the risk of confounding. What is more, along with a thicker conception of sex comes dispute about what exactly constitutes sex, undercutting interventionism's claim to specifying a unique counterfactual and opening the door to substantive moral and political judgments. When operationalized in a test for discrimination, interventionism paired with an unsettled conception of sex generates ambiguity about how to actually run the test. These reasons press towards adopting a thin conception of sex status that distinguishes it from other social factors that may be non-accidentally correlated with sex but are not sex "itself."

All that being said, I am less inclined to insist here that a "true" interventionist analysis of discrimination *must* adopt a thin conception of sex (though I do take a more insistent tone in the previous chapter) and set aside the matter of whether an interventionist could in theory formulate a different, better causal test for discrimination. My target here is the interventionist conception of

[107] Catherine A. MacKinnon, "Difference and Dominance: On Sex Discrimination" in *Feminism Unmodified: Discourses on Life and Law* (Cambridge: Harvard University Press, 1984), 32–45.

discrimination which predominates in legal practice, which as a matter of fact does take sex's causal effect to be distinct and separable from those of other factors. An inquiry to the causal significance of sex to an employer's action must therefore, per the test, take care not to launder in these other effects. The interventionist matches sex contrasts along as many causally significant factors as possible, stripping out by the method of elimination more and more of their potentially confounding effects, so that sex can finally have its causal effect once "everything else" has had theirs. The interventionist's substantive conceptual assumption rules that differences in compensation across men and women that are chalked up to, say, differences in the kinds of positions predominantly occupied by women (e.g., cleaning, child care, domestic service) compared to the kinds of positions predominantly occupied by men (e.g., manufacturing, craftsmanship, technical expertise) may not be caused by "sex" but rather differences in market valuation of different types of labor and skills. And so on and so forth. The interventionist's social ontological error thus yields errors in her causal judgments, which, in its deflationary conclusions about the causal significance of sex and race, serves to reinforce the mistaken view that sexes and races are, as social groupings, fundamentally the same or interchangeable.

Of course, there is no view from nowhere, and the interventionist has not in fact eschewed all normativity in adopting the thin account of sex. Her analysis of sex's causal significance, which separates out sex from that which it systematically correlates, deflates sex into a status that has little to no social reality. I asked in the previous chapter and here I ask once again, what the effect that remains at the end of such an eliminative procedure is an effect *of*? For the purification process would have seemed to erode any basis for the category to have any causal effects at all. I suggested that the "just sex" residue that remains can be little more than the effect of the physical traits themselves or the effects of a pure affective (dis)taste for these traits. If these are the whole of effects ruled out by prohibitions on sex discrimination, then a mandate of nondiscrimination calls

for little more than neutrality with respect to the trait itself. Now if sex *were* a status with little social footprint, such an account of nondiscrimination with respect to sex would be perhaps harmless albeit peculiar, for it would seem to be an embargo on what amounts to a kind of irrationality. We might similarly forbid decisions from being influenced by how an individual ties their shoelaces or whether they prefer their fries curly or straight. But that sex and race discrimination is, as most agree, an act of greater moral and political significance than mere idiosyncratic decision-making based on shoelace-tying technique and fries preference is rationalized by the fact that the social categories of sex and race are, again as most agree, categories defined by social difference. They are categories characterized by differences in the distribution of social benefits and burdens on the basis of certain phenotypic features, where differences in these "non-sex" or "non-race" factors are not distinct from sex or race "itself"; rather, they are constitutive of what it is to be sexed or raced.[108] Hence, the interventionist causal conception of discrimination delimits employers' responsibility for these broad racial and gender inequalities and releases them from any obligations that might derive from them. The resulting account of nondiscrimination as neutrality neglects these background social facts, permitting—and perhaps even *demanding*—that employers neglect them, too.

§2.2 The Epistemic Assumption

An interventionist who pursues causal inquiry in practice must assume she has knowledge of the full set of time *t* factors relevant for her outcome of interest. Only then may she proceed by neutralizing their effects one by one per the method of elimination that underwrites her causal test. Dispute about whether this condition is satisfied or is justified weakens any causal conclusions she might go on to draw by opening them up to methodological critique. A skeptic of some factor's

[108] On accounts of sex and race that emphasize their being defined by these social differences, see e.g., Sally Haslanger, *Resisting Reality* (Oxford: Oxford University Press 2012); Esa Díaz-León, "What is Social Construction?," *European Journal of Philosophy* 23, no. 4 (2015): 1137–1152.

causal efficacy can always charge that under a certain interventionist analysis, the factor only *seems* to have an effect because the true cause has been omitted from the analysis. And had it been included for consideration, the analysis would have found that the factor in question makes no difference after all. By the same token, the specter of unaccounted for causal confounders can also be marshaled to call into question the finding of a *lack* of causal effect. When the epistemic assumption is in question, causal conclusions are characterized by epistemic instability and a baseline level of tentativeness and uncertainty.

The arguments in *Sears* are here illustrative. To support their claim that Sears engaged in nationwide discriminatory practices on the basis of sex, EEOC offered statistical analysis showing that controlling for the job applied for, age, education, job type experience, product line experience, and commission product experience, those sexed female were significantly less likely to be hired and promoted at Sears. The substantial disparities that remained even after making adjustments to account for the influence of these other variables, EEOC argued, showed the causal relevance of sex to the company's decision-making and hence, sex discrimination in hiring and compensation.

An interventionist causal reasoner is persuaded to the extent she believes the epistemic assumption to hold: that the six factors controlled for in EEOC's analysis exhausts the set of causally relevant factors for hiring and promotion decisions at Sears. And unsurprisingly, Sears was not so easily convinced. Their rebuttal held that EEOC's analysis failed to account for *all* factors relevant to hiring, promotion, and compensation decisions at the company. Putting forth statistics-based causal inference analyses of their own, Sears showed that upon controlling also for "veteran status, marital status and size of family, leaves of absence and college major," the effect of sex diminishes by an average of more than 62%.[109] Moreover, claimed Sears, EEOC's neglecting to rule out the causal relevance of admittedly more difficult to measure but still important factors such as

[109] *EEOC v.* Sears, 628 F. Supp. at 1344, 1345.

"physical appearance, assertiveness, the ability to communicate, friendliness, and economic motivation" further undercuts their case for having homed in on the effect of just sex. The plaintiff's analysis was as good as "comparing apples to oranges."[110]

I want to note two things about this reply. First, the response suggests that showing that hiring and compensation decisions were made on the basis of those factors that Sears here offers would work to successfully rebut the charge of discrimination on the basis of sex. That is, it assumes that decisions made on the basis of *any* factors at all so long as they are not sex status *per se* is admissible as nondiscriminatory. But this presumption seems to me too quick. Why should, for example, compensation based on veteran status necessarily release Sears from any claim of sex discrimination, given the relationship between sex status and military service? It seems to me that factors which are closely related to sex may fall under scrutiny as well. This point harkens back to the earlier discussion about whether the sharp distinction the interventionist draws between sex and those factors which constitute sex as a social status in her analysis is really tenable.

The second is a point about *strategy* in the tug-of-war for causal proof. If, granting the interventionist's distinctions, decisions on the basis of any factors that are not sex "itself" do absolve the accused of charges of discrimination, one might wonder why the defense does not always cite whichever factors that will show the outcome to be less dependent on the factor that is sex. Indeed as it turns out, given this interventionist causal standard, defendants in discrimination trials typically *do* respond just as Sears does here: by charging that the opposing side's model suffers from "omitted variables," factors which were in fact causally relevant to the outcome—or at least, so they claim—and whose absence generates an artificial effect of the status at issue. In support of this claim, they

[110] *EEOC v.* Sears, 628 F. Supp. at 1290; "Trial Brief of Sears, Roebuck, and Co." at 9. It bears noting that EEOC disputed that proper analysis of the effect of sex should include the variables Sears proposes, on the grounds they lack explanatory power. The court here sided with Sears, arguing that even when variables' explanatory effects are uncertain, they should nonetheless be included in the model. Ruth Milkman "Women's History and the Sears Case," *Feminist Studies* 12, no. 2 (1986): 375–400.

are incentivized to put forth a "kitchen sink" causal analysis containing any and all factors that an employer may legally consider in employment decisions, that results in a deflated causal effect of the protected attribute at issue. And since plaintiffs are hard-pressed to substantiate objections claiming that employment decisions did not in fact draw on such considerations, defendants in practice enjoy substantial leeway in producing an analysis in this vein.

Crucially, defendants are not only incentivized to reply with "kitchen sink" analyses so to successfully escape legal liability. Interventionism as a theory of causal epistemology *itself* prefers such analyses. For as far as the method of elimination is concerned, an analysis that accounts for more variables and thereby eliminates too their causal contributions is strictly better than a sparser model at homing in on the effect of sex. An analysis that looks to pin causal influence on, say, sex *should* as a matter of methodological hygiene include as many plausible controls as possible, lest it, as Sears charged, compares "apples to oranges."[111] Recognizing this, the *Sears* court wrote in their opinion siding decisively with the defendant:

> EEOC's statistical analyses lack persuasive value because EEOC omitted important variables which influence checklist compensation at Sears. It is important to bear in mind that, with the type of statistical model used by both EEOC and Sears, all important factors which affect compensation must be included in the model to obtain reliable results. The analyses produce a sex coefficient, which is intended to measure the effect of sex on compensation. However, the sex coefficient reflects not only the effect of sex, but also the residual effect of any factor which affects salary that is not included in the model. Thus, if important variables are omitted, the effect of sex on compensation estimated by the model will be artificially inflated… [I]n this situation, it is better to err on the side of including variables when their effect on salary is uncertain, than to exclude them and obtain a biased estimate of the sex effect.[112]

[111] In actual cases of discrimination litigation, there is disagreement about the extent to which variables that might themselves be downstream from discrimination may be permitted in an econometric analysis. For example, an employer might look to fend off a charge of sex discrimination in compensation by arguing that women were less productive on the job than were men and thus showing that accounting for productivity levels resolved the wage gap. A court may rule that the analysis may not be allowed to enter as evidence in the case on the grounds that the difference in productivity is itself due to sex discrimination in, say, the resources that the employer offers to men vs. women. This can be read as a challenge to the time at which the intervention is allowed to swoop in to alter sex status in generating the right counterfactual contrast.

[112] *EEOC v.* Sears, 628 F. Supp. at 1344.

Although the court is here mistaken about the *direction* of bias that omitted variables generate—it is not the case that omitted variables always positively bias a measured effect[113]—its preference for the defendant's expanded model is well justified by lights of good interventionist causal reasoning. Including variables, even those which are irrelevant to employer's decisions, presents no risk of biasing the effect of interest, but excluding relevant variables certainly does. Hence, on the use of statistical tests in discrimination litigation, legal scholar and economist Ian Ayres writes that the "'kitchen sink' approach… [is] the standard method of testing for disparate treatment discrimination" with the theory behind it "well settled."[114]

Of course, the particular litigation dynamics, statistical analyses, and ultimate judgment in *Sears* cannot stand in for the general matter of how the interventionist causal test tips the scales in legal analyses of discrimination. What matters for my argument is that the positions staked out by both EEOC and Sears, as well as the court's reasoning across them is largely in keeping with the core of interventionist causal thinking. In other words, the ruling in *Sears* was not simply the product of an ideologically-motivated interpretation of statistical analysis; rather it was the upshot of interventionism's troubled approach to reasoning about sex as a cause. The court's tendency towards accepting defenses against charges of discrimination makes for an onerous standard of proof for those who are claimed victims of discrimination. And although this bias is inherited from the

---

[113] Whether omitted variable bias generates an under- or over-estimate of the effect of some variable depends on the correlations between the omitted variable and the target variable of interest, on one hand, and the omitted variable and outcome, on the other. If the omitted variable shares the same direction of correlation with both the target variable and outcome variable, then its omission in the model results in an overestimate, or upward bias, of the effect of the variable of interest. If the omitted variable is negatively correlated with the target variable and positively correlated with the outcome variable, or vice versa—i.e., if the pair of correlations are in opposite directions—then its omission in the model results in an underestimate, or downward bias, of the effect of the variable of interest.

[114] Ian Ayres, "Testing for Discrimination and the Problem of 'Included Variable Bias'," working paper, 2010, 13. It is important to note, as Ayres argues here, that the kitchen sink approach should not be taken to be best practice in discrimination trials more broadly, since an employer may be found guilty of discrimination via the disparate impact doctrine, under which adverse outcomes against a protected class may constitute discrimination in themselves unless justified by reference to legitimate business purposes. In these cases, a kitchen sink regression may suffer from "included variable bias," if variables that are not so justified are erroneously allowed to enter the analysis and deflate the causal effect of the protected trait.

conditions of soundness of the method of elimination which itself lacks moral content, here it has a distinctively political character.

I have argued that a theory of discrimination which takes from an interventionist conception of sex causation adopts a premise of sex sameness, fills in the content of nondiscrimination on the basis of sex (or sex equality) as sex neutrality, deflates the causal significance of sex in the social world, and is, in practice, biased against findings of sex discrimination. These normative upshots share a distinctively politically conservative bent. If it turns out that sex and race play a minimal role in determining individuals' life outcomes, interventions targeting these social categories as a site of social (dis)advantage are either misled and unnecessary or at least not supported by the evidence. And if one takes no issue with the mere fact of inequalities, these conclusions support preserving a domain of private actor and market decision-making free from state intervention.

It might strike one as somewhat surprising that a test for causation could supply a theory of discrimination with substantial normative content. For whether something is or is not causally relevant to some outcome seems to most philosophers to be a wholly non-normative matter. I set forth a challenge for this popular view in the preceding chapter—at least as it pertains to the interventionist. Here my argument that an account of discrimination outfit with a particular interventionist picture of causation acquires alongside the causal standard, substantive moral and political positions can be taken as another clue that suggests interventionist analyses of categories such as race and sex to give value-laden accounts of social causation.

But there is more to the interventionist causal standard of discrimination than these first-order normatively troubling upshots. That the analysis is presumed to be innocent of ethical content generates further, markedly political problems. It is to these pernicious second-order effects of the interventionist discrimination analysis that I now turn.

§3. Second-order normative upshots of the interventionist conception

In the preceding section, I argued that the operationalization of interventionist reasoning into a legal test for causation in discrimination trials fills in substantive content for the legal and ethical concept, despite its seeming to wholly lack moral and political content. Verdicts about sex and race causation, and by extension about sex and race discrimination, appear as plain announcements of natural, non-normative causal facts. All the while, the interventionist causal test's political character when applied to social categories such as sex and race is hidden from view. I want to now draw out two consequences that follow from these false appearances.

When figuring determinations of discrimination appears as a straightforward empirical task and a matter of sorting out a set of non-normative causal facts, the verdicts issued appear objective, in the sense of being independent of moral and political considerations and so incontestable by them. Someone who responds to a claim that sex did not cause an adverse outcome with a moral objection would seem to be committing a rudimentary category error. After all, she cannot contest a deliverance of descriptive fact about what *is* the case with a normative claim about what *ought* to be. Interventionist analysis takes a question of whether a given state of affairs counts as discrimination for a question of whether the relevant causal facts do or do not obtain per the interventionist causal test. Shifting debate about discrimination from the space of reasons to the space of causes works not only to short-circuit ethical reasoning. It exerts also a discourse-shaping effect on how we reason and dispute different accounts of discrimination. In this regime, debate over what constitutes discrimination no longer involves making moves in normative space from normative standpoints. Instead, disagreement over cases reduces to empirical dispute about causes and effects and whether causal conclusions drawn are epistemically sound given the accepted means of doing causal analysis.

I think it is helpful at this point to return to the early years of the EEOC's enforcement of the then new discrimination protections and remind ourselves how far the causal inference call-and-response approach to disputes about discrimination that prevails today is from the claims of discrimination put forth by the thousands of women who wrote into the agency and from the agency's own strategy in pursuing and resolving such cases. Claimants were not making implicit interventionist causal arguments about sex status and the challenges they faced at work, nor is it fruitful (or accurate) to say they meant to. Working women's concerns often stemmed from an explicit acknowledgment of sex *difference*.[115] "Comparable worth" campaigns were centrally concerned with the intrinsic value of femininized labor—labor, which by definition has no male analogue. Its advocates argued that secretarial labor, waged social reproduction, domestic work overwhelmingly performed by women was intrinsically valuable to a workplace's functioning and ought to be valued by employers as such. In the struggle for fair pay for the differently sex-typed housekeeping positions of maids and housemen in hotels, working women openly acknowledged that labor performed by the former and the skills required—performing repetitive tasks under great time pressure, being in close contact with guests, being subject to higher standards of accountability—was *not* interchangeable with, or essentially the same as, that of the latter and their corresponding skills—moving furniture as needed, fixing appliances, cleaning spaces outside of guestrooms.[116] Still, maids argued that the different housekeeping positions were comparable in their *intrinsic worth* to the hotel's functioning nonetheless, their *market valuations* notwithstanding. Untroubled by the possibility that it might undercut their claims of discrimination, working women argued for the relevance of sex difference in determining what constitutes equal treatment on the

[115] As Katherine Turk chronicles in *Equality on Trial*, many working women opposed systems that forced them to take male-typed jobs and many who wrote into the EEOC argued that their sex should explicitly taken into account in determining what their benefits, rights, and protections should be in the workplace. Turk, *Equality on Trial.*
[116] Turk, *Equality on Trial*, Ch. 5.

basis of sex in the workplace. The grounds of their claims were normative through and through: that the new law *ought* to apply to their case, that their situation *ought* to be considered sex discrimination, that women's work *ought* to be valued *in spite* of its differences to men's work.[117]

It is always available, of course, to respond by saying that however unfairly treated, many of these women were simply misled about being victims of employer discrimination and were wrong about the meaning of Title VII. This objection, however, first neglects to acknowledge that many of these women did win their claims to workplace sex equality at the time, and many employers did accede to the EEOC's demands for repair. No doubt that they did so on an entirely different basis than would pass legal scrutiny today. But to claim that the "right" analysis of discrimination is the one that in the end won out has more than a whiff of whiggism. It is to read the past as mere precursor to the present, to accept as "right all along" whatever happens to be the case today. This is certainly bad historiography for one, worse still as an orientation towards theorizing our normative concepts. Here it not only begs the question in favor of the interventionist conception of discrimination but any (purportedly) amoral descriptive causal analysis.[118]

The objection looks to reroute debate back towards interventionist terrain. But as comparable worth campaigns of the 70s and 80s show, many of the early substantive conflicts at issue in the early years of discrimination law resist causal reasoning entirely. Does the wage gap resulting from the labor market's devaluation of feminized work constitute discrimination on the basis of sex? Can jobs *across* the gendered division of labor be scrutinized for sex discrimination? Given gendered differences in the distribution of the burdens of social reproduction outside of the sphere of paid work, should employers be able to offer high wage premiums to employees able to take on long and inflexible hours? These questions were "live" in a previous historical moment and

[117] Ibid.
[118] Thanks to Liam Bright for this suggestion.

nothing about the concept of discrimination seems to preclude them.[119] It is only by interventionist causal standards that these debates cannot get off the ground. By setting down a substantive view of what discrimination is, the interventionist causal standard therefore also sets the agenda on what proper debate over discrimination looks like. The right comparisons to make to adjudicate matters of discrimination, so says the interventionist discrimination theorist, are apples-to-apples comparisons premised on sameness across sexes and races. The right kind of objection to another's determination of discrimination is one that aims at undercutting their causal analysis. If you wish to engage on the terrain of discrimination, you must engage on the terrain of interventionist causal reasoning. To the extent that these questions cannot conform to these rules of engagement, they are either not bona fide cases of discrimination or do not fall under the purview of the concept at all.

Setting the agenda in this vein manages a version of what Toni Morrison calls the "distraction" function of racism, in this case operating to continually shift the burden onto those who are subordinated to prove the existence of discrimination according to this dominant conception of what constitutes proof.[120] An onerous burden of proof not only minimizes a claimant's chances of successfully substantiating a charge of discrimination, efforts at meeting the standard are highly taxing. Those who claim discrimination must engage in complex sparring of causal evidence, methodically eliminating all non-discriminatory alternative explanations that their opponents may proffer as the "true" cause of the disadvantageous outcome. Hence, discourse that centers proper causal methodology directs energy and attention away from the task of challenging

[119] Although in *County of Washington v. Gunther*, the Supreme Court distinguished comparable worth claims from other cases of intentional discrimination, it declined to rule whether a comparable worth argument could on its own establish a case of discrimination prohibited by Title VII. County of Washington v. Gunther 452 U.S. 161 (1981).
[120] Toni Morrison, "A Humanist View," Oregon Public Speakers Collection: Black Studies Center Public Dialogue. Pt. 2, Portland State University, May 30, 1975.

on normative grounds prevailing conceptions of discrimination and putting forth alternatives. As Morrison says, "It keeps you from doing your work."[121]

In this tendency towards misdirection to the effect of bolstering an unjust prevailing social order, a theory of discrimination equipped with interventionist causal reasoning can be considered in part *ideological,* along the lines of Charles Mills's critique of "ideal theory" as ideology.[122] For Mills, concepts and constructs of ideal theory are "patently deficient, clearly counterfactual and counterproductive approach[es] to issues of right and wrong, justice and injustice."[123] And yet these mistaken "distortional complex of ideas, values, norms, and beliefs" prevail as the dominant mode of theorizing in political philosophy because they reflect and serve the "nonrepresentative interests and experiences" of the bourgeois white males who dominate the professional discipline of philosophy.[124] He writes:

> [A]ll theorizing, both moral and nonmoral, takes place in an intellectual realm dominated by concepts, assumptions, norms, values, and framing perspectives that reflect the experience and group interests of the privileged group (whether the bourgeoisie, or men, or whites). So a simple empiricism will not work as a cognitive strategy; one has to be self-conscious about the concepts that "spontaneously" occur to one, since many of these concepts will not arise naturally but as the result of social structures and hegemonic ideational patterns. In particular, it will often be the case that dominant concepts will obscure certain crucial realities, blocking them from sight, or naturalizing them, while on the other hand, concepts necessary for accurately mapping these realities will be absent.[125]

---

[121] Morrison, "A Humanist View."
[122] Charles Mills, "'Ideal Theory' as Ideology," *Hypatia* 20, no. 3 (2005): 165–184.
[123] Mills, "'Ideal Theory' as Ideology," 172.
[124] Ibid.
[125] Ibid., 175

While Mills's critique aims at the overly broad idealizations of ideal theory in political philosophy, my charge against the interventionist causal conception of discrimination is that the account is overly narrow and does not correctly conceive of the encompassing nature of sex and race categories. These might appear at first to be opposing diagnoses, but they are in fact two sides of the same coin. Just as for Mills, idealization in ideal theory serves to paper over existing injustices, the premise of sex and race sameness that underwrites the interventionist's test for discrimination similarly obscures the reality of racial and sex hierarchy. What is scrutinized in both is the starting points of ethical theorizing that stand clearly contrary to social fact and contrary to our end of theorizing concepts toward ethical improvement in practice.

If this is right, then why, Mill presses us to ask, are they nevertheless well accepted as adequate premises of our moral, political, and legal theorizing? On the suggestion that the account is ideological, the answer is that theorizing in this vein reflects the position of those whose experience navigating the social world does not centrally involve taking heed of the hierarchical structures of sex and race.[126] Meanwhile the resulting account of nondiscrimination as neutrality aligns with the material interests of those groups who would bear the costs of redistribution. Even while the judgments of the deflated footprint of racial and sex discrimination are naturalized and look to be reflections of how the social world just is, they in fact emerge from the particular viewpoints of dominant classes and function to preserve the social order in their interest.

I want to suggest now that the ideological effects of this analysis of discrimination and causation are more pernicious than is indicated by this account in Mills alone. The worry about ideal theory as ideology is that if our conceptual frameworks and toolkits shape how and what we think, then bad political philosophy will eventually yield bad political practice. While I do not doubt that

[126] See e.g., Patricia J. Williams, "Alchemical notes: Reconstructing ideals from deconstructed rights," *Harvard Civil Rights-Civil Liberties Law Review* 22, (1987): 401.

there is a link between what we theorize and what we do, I am often uncertain as to the precise trickle-down mechanism by which discourse reaches out to touch the goings-on of the "real" world. However, in the case of the interventionist conception of causation and discrimination, the link between the analytical framework and material outcomes appears to me wholly demystified. For the agenda-setting capacity of the interventionist approach to causal thinking about race and sex is not contained within only the sphere of philosophical discourse. It rules also in the social sciences, dictating the standards of rigor required for causal claims to be deemed properly "scientific." Social scientific inquiry that aims at the interventionist "gold standard" in its study of race and sex finds their causal significance diminished. When the institution of science is taken to be a society's best arbiter of matters of cause and effect, it is easy to see how theoretical debate about sex and race causation and discrimination shapes the space of social and political responses to raced and gendered features of our social world. The dominant paradigm of causation and discrimination forms scientific methods and verdicts which in turn plug in to key social institutions, facilitating the development of a set of legal evidentiary standards, policymaking guidelines, and debate among experts that bear profoundly on our collective capacities to diagnose and respond to social harms and injustices. These are the means by which the first-order normative upshots I discussed in §2—comparisons premised on an assumption of race and sex sameness, deflated causal estimates, a strong presumption against the existence of discrimination—descend from the realm of theoretical possibility to shape social and material life. The shift from normative debate to debate about causal effects thereby marks also a significant transformation of the deliberative process by which a polity figures the social and ethical question of what constitutes discrimination. In this form, discrimination becomes a matter of specialized technical knowledge for which scientific expertise is necessary. It becomes a domain within which econometricians, apparently, enjoy an authoritative

voice and ordinary people, meanwhile, ought to defer to those better trained in causal inference before drawing ethical conclusions.

§4. Conclusion

And yet it is clear that the value-laden nature of the problem of discrimination cannot truly be dissolved, no matter the sophistication of attempts at hunting causes and effects. My claim throughout this chapter has been that, contrary to appearances, not even the interventionist causal conception manages to *displace* the normative content of an analysis of discrimination. It rather *fills out* this substantive content in its causal standard, which has the effect of ruling that only a narrow band of cases are even candidates for triggering concerns of discrimination. The interventionism-based legal test therefore traces out a tight boundary of responsibility that largely relieves employers of the need to consider how their practices may differentially benefit and burden individuals on the basis of race and sex against a background social structure that is ordered by race and sex. What employers owe to workers is equal consideration, independent of these facts. On the analysis that presently dominates, antidiscrimination doctrine is a legal instrument that secures only a thin formal equality of opportunity and is a weak guarantor of justice for those marked by social statuses of disadvantage.

Some theorists of discrimination seem to agree with all that I've so far said, and simply take these normative upshots to be unfortunate facts about the concept of discrimination.[127] The thought here is that boundaries of the interventionist conception of discrimination are drawn more or less accurately, regrettable this fact may well be. I find such a response puzzling, at least as a reply from the theorist engaged in conceptual analysis of a moral or political notion. For the central question

[127] See e.g., Benjamin Eidelson, *Discrimination and Disrespect*.

before us in any of our endeavors that subject ethical terms to scrutiny is precisely whether we *ought* to adopt a given analysis. On the assumption that a good account of discrimination, the moral concept, will draw on and also look to inform our actual social and legal purposes and practices pertaining to discrimination, one of our key questions here is: what legal standard does best to safeguard those interests that prohibitions of discrimination are formulated to secure? It seems strange to admit on one hand that the predominant analysis renders antidiscrimination as a legal instrument inadequate vis-à-vis its functional role in, say, securing workplace sex equality, while on the other, accede to it as the "right" conception regardless. One takes on these bad normative upshots only if one is already pre-committed to the winnowed down conceptual space of discrimination that presently prevails. Though, of course, to adopt this view is to simply beg the question of whether we *should* retain the interventionist causal analysis.

Taking a historical lens to the legal notion does well to prevent this mistake, as it reminds us that presently drawn boundaries of some concept are by no means pre-ordained. In the case of Title VII's sex provision, a historical treatment sees the moral concept of discrimination as a living entity that took shape over time via competing claims put forth by workers, employers, bureaucrats, judges, legislators, activists, and so on, who were motivated by their real, often immediate, material interests in the conception of sex discrimination that would win out, that is, on what constitutes fair or unfair treatment *because of* or *on account of* sex. A historical perspective thus lends itself to a broadly *materialist* approach to conceptual analysis, which takes the content of discrimination to be formed out of the efforts of agents doing things in the world to stretch and pull and shrink and compress a notion about which they had significant reason to pursue change. Those who fought for equal pay for female-typed and male-typed housekeeping work argued that hotel compensation schemes that paid maids less than housemen treated women unfairly and thus discriminated on the basis of sex. Management, unsurprisingly, disagreed. Looking to avoid owing millions in backpay to women

laboring under a discriminatory contract, hotels fought back against the charges by arguing that maid and housemen positions were different in more than just sex type. Unequal remuneration across them did not constitute unfair treatment because of *sex*, so they claimed, because the positions were different in *other* ways: in their respective duties and tasks, their difficulty, and market valuations. As they saw it, these differences justified differences in compensation, and so lower wages for maids did not constitute sex discrimination in pay. This is a dispute about what sex discrimination *is*, what interests are meant to be secured by the new law, and what ensuring sex equality in the workplace demands of employers. Working women argued that laws against sex discrimination ought to ensure that feminized labor would be valued and remunerated accordingly by employers precisely because of its long-standing unjust devaluation in the market and society more broadly. Employers sought to preserve their autonomy over pay-setting, arguing that only gaps in compensation that were incontrovertibly because of sex status "itself" warranted state intervention. Naturally, this meant favoring a strict standard for what counts as "because of sex" and, as a result, a highly restricted range of cases that would draw legal scrutiny and warrant incursions by the state. Today's leading interventionism-based analysis certainly delivers these victories for employers, but this fact by no means vindicates their substantive position as giving the right conception of discrimination.

When offered up as an analysis of discrimination, the interventionist account of what it is for sex and race to be a cause fills in also an account of what constitutes unfair treatment on the basis of sex and race. My aim in this chapter has been to show the political character of the legal standard, and in this dissertation more broadly, the normative content of interventionist reasoning about sex and race. One payoff of this argument is that it undercuts any claim to legitimacy that the prevailing conception of discrimination might earn on the back of its claim to operationalizing a supposedly non-value-laden analysis of causation. Left with no non-normative ways out via analyses of causation, we are forced to face up to the ethical core of the discrimination question: what

constitutes fair treatment on the basis of categories of sex and race? Reorientation around the fact of racial and gender inequality exposes clear deficiencies in the fairness standard set by the interventionist causal conception. An analysis of discrimination that cannot register any raced and sexed differences in employment outcomes that elude apples-to-apples comparisons both reflects and reinforces existing raced and sexed injustices. What is more, the starker these disparities, the more deep-seated the inequalities, the less the analysis is able to speak to it, and the less a prohibition of discrimination is able to secure for those in positions of social disadvantage. At the extreme, a society characterized by hierarchy so calcified such that no two individuals from different strata are sufficiently alike to be appropriate comparators rules out discrimination by definition. Antidiscrimination as a legal instrument becomes utterly futile; and those in subordinate strata can make no claims to being treated unfairly on account of their status at all.

If only the point were a far-off hypothetical. Sadly, the exact issue has played a hand in curbing the reach of antidiscrimination law in fact. As legal scholar and sociologist Issa Kohler-Hausmann has pointed out, a history of racial segregation has made many neighborhoods in the US incomparable in precisely the ways that render it impossible to meet the interventionist causal standard.[128] Such neighborhoods are simply too different across too many other relevant features to warrant drawing conclusions about the causal significance of race in particular, stymying the ability of claimed victims to successfully bring forth charges of racial discrimination. Discrimination cases regarding sex-typed work have drawn out similar arguments and conclusions. Even as explicitly sex-segregated positions were eliminated, employers continued to pay less for female-typed work, arguing that feminized labor itself was less difficult, in less demand, or for whatever other reasons,

[128] Issa Kohler-Hausmann, "Eddie Murphy and the Dangers of Causal Counterfactual Thinking About Detecting Racial Discrimination," *Northwestern University Law Review* 113, no. 5 (2019): 1163–1228, 1218. For the statistical problem that lack of overlap poses see Robert Barsky, John Bound, Kerwin Ko' Charles, and Joseph P. Lupton, "Accounting for the black-white wealth gap," *Journal of the American Statistical Association* 97, no. 459 (2002): 663–673, on the fact of little overlap between the Black and white earnings distribution.

simply worth less. Employers claimed the existence of unbridgeable differences in female-typed and male-typed work to beat back working women's demands for higher pay. Without a sufficiently "similar" position that was better waged, women in female-typed positions have no claim at all as to whether their pay is or is not discriminatory.

In the case of sex-typed work, the interventionist causal conception does not just supply employers with a reliable defense against charges of discrimination, it has proven to be a formidable tool in a positive project of constructing a theory of discrimination that aligns with their material interests. Nondiscrimination might well demand equal remuneration for male-typed and female-typed work that is interchangeable, but it makes no claims as to whether what is morally required to resolve instances of sex discrimination is that women's wages be raised up to the level of men's. This leaves employers with the option of leveling-down: restructuring their labor force, eliminating positions, and reorganizing, combining, and intensifying duties so that positions that remain are indeed more interchangeable, such that all workers, men and women alike, are paid, equally, at the same meager wages. And indeed, over the course of the countless drawn-out legal battles over equal pay that took place in the 70s, 80s, and 90s, this is precisely what happened in many industries. To return to the decades-long struggle over the wages of male and female-typed housekeeping positions, even while hotels argued that the positions of maids and housemen were fundamentally incomparable such that maids' lower pay did not constitute sex discrimination, they pursued longer-term labor restructuring strategies to make the two housekeeping roles increasingly indistinguishable. By the 1990s, hotels had largely agreed to eliminate sex-based distinctions in their workforce, phasing out the housemen's position entirely and reorganizing their labor force into housekeeping roles that were ostensibly gender-neutral but with substantially more burdensome duties and task quotas—all in the name of abiding by the new sex discrimination law.[129]

[129] Turk, *Equality on Trial*, Ch. 5

I mention these historical details to stress what the *stakes* of engineering a concept like discrimination are. The passage of Title VII enshrined new rights, and so it enshrined new claims that people could make on others and on the state, such that when they are harmed or wronged, they have claims to compensation or restitution. Claim-rights entail correlate duties that are owed and costs that must be incurred. Employers and whoever else are liable to face up charges of discrimination do not only have an interest in not being saddled with these duties and costs; that is, they do not only have an interest in a narrow conception of what counts as discriminatory so that they are obligated by a narrow set of duties. They have an interest in reconfiguring the concept, such that even the correlate duties of antidiscrimination protections are aligned with their interests. Corporations that may benefit from eliminating sex-segmented work in their labor forces are not only ruled to be in good legal standing in doing so but have their choices fortified by antidiscrimination law. Similarly, employers under fire for racially homogenous work forces appeal to principles of colorblindness as a demand of nondiscrimination. Agents who are typically in the position of the duty-bound vis-à-vis discrimination law have interests not only in avoiding bearing the costs of securing the moral goods that nondiscrimination is meant to pick out, but to foreground a different set of moral goods entirely, protection of which they too can benefit from. These have always been the real stakes of "conceptual engineering" the notion of discrimination in the real world.

These observations lend themselves to something of a methodological point about our philosophical theorizing about ethical concepts. Many of the values and criteria that philosophers hold dear in their analysis of some ethical concept—adherence with our intuitions, delivery of the proper verdicts on the supposed "clear-cut" cases, consistency with the best accounts of related concepts, non-overlap with nearby ethical notions, and so on—are largely untethered to the realm of earthly affairs. On the dominant way of doing things, the philosopher starts from prevailing

conception(s), weighing them against each other according to these virtues, tinkering where needed, and finally smoothing out edges by resolving ambiguities that yet remain at the margins. A materialist perspective, by contrast, sees that the content of our concepts is formed out of the efforts of agents doing things in the world to stretch and pull and shrink and compress a notion about which they have significant reason to pursue change. On this view, it is people's material interests in what discrimination as a moral and legal concept can *do*, rather than considerations of conceptual clarity or theoretical consistency, that weighs most heavily in favor of one or another analysis.

One aim of this chapter has been to argue that a dominant account of discrimination, one that presently constitutes the legal standard, yields moral and political upshots that show it to be poorly suited to the task of securing those moral goods that the concept is formulated to ensure. Located centrally in that argument are the practical stakes of theorizing discrimination, the historical development of the predominant analysis, and the practical upshots that have followed from it. Hence I see this chapter also as attempting a materialist conceptual analysis, built around the undeniable fact that people have real stakes in how conceptual engineering of a notion of like discrimination goes. Thus a second, more oblique, aim of this chapter has been to defend a place in our philosophical theorizing for this materialist approach. For in the real world, for better or for worse, an account's "winning out" is a matter wholly orthogonal to the philosopher's standards of "getting it right" and discrimination is little more than these practical upshots. In the real *unjust* world, we limn the prevailing concept at our peril.

Bibliography

Agan, Amanda, and Sonja Starr. "The Effect of Criminal Records on Access to Employment." *American Economic Review* 107, no. 5 (2017): 560–564.

Ayres, Ian. "Fair driving: Gender and race discrimination in retail car negotiations." *Harvard Law Review* 104, no. 4 (1991): 817–872.

———. "Further evidence of discrimination in new car negotiations and estimates of its cause." *Michigan Law Review* 94, no. 1 (1995): 109–147.

———. "Testing for Discrimination and the Problem of 'Included Variable Bias'." working paper, 2010, 13.

Barsky, Robert, John Bound, Kerwin Ko' Charles, and Joseph P. Lupton. "Accounting for the black-white wealth gap." *Journal of the American Statistical Association* 97, no. 459 (2002): 663–673.

Baumgartner, Michael. "Interventionist Causal Exclusion and Non-Reductive Physicalism," *International Studies in the Philosophy of Science* 23, no. 2 (2009): 161–178.

———. "Interventionism and Epiphenomenalism." *Canadian Journal of Philosophy* 40, no. 3 (2010): 359–384

———. "Rendering Interventionism and Non-Reductive Physicalism Compatible." *dialectica* 67, no. 1 (2013): 1–27.

Bertrand, Marianne, and Esther Duflo. "Field Experiments on Discrimination." In *Handbook of economic field experiments Vol. 1*, edited by Abhijit Vinayak Banerjee and Esther Duflo. Amsterdam: North-Holland, 2017, 309–393.

Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg More Employable Than Lakisha and Jamal: A Field Experiment on Labor Market Discrimination." *American Economic Review* 94, no. 4 (2004): 991–1013.

Blank, Rebecca M., Marilyn Dabady, and Constance F. Citro. *Measuring Racial Discrimination*. Washington, DC: The National Academies Press, 2004.

Brooks, David. "What is discrimination discrimination?." *Philosophical Papers* 11, no. 1 (1982): 15–30

Burge, Tyler. "Individuation and Causation in Psychology." *Pacific Philosophical Quarterly* 70, no. 4 (1989): 303-322.

Campbell, John. "An Interventionist Approach to Causation in Psychology." *Causal Learning: Psychology, Philosophy, and Computation* (2007): 58-66.

———. "Independence of Variables in Mental Causation." *Philosophical Issues* 20 (2010): 64-79.

———. "Interventionism, Control Variables and Causation in the Qualitative World." *Philosophical Issues* 18 (2008): 426-445.

Dembroff, Robin, and Issa Kohler-Hausmann. "Supreme Confusion About Causality At the Supreme Court," *CUNY Law Review* 25, no. 1 (2022): 57–92.

Deming, David J., Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz. "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study," *American Economic Review* 106, no. 3 (2016): 778–806.

Díaz-León, Esa. "Substantive metaphysical debates about gender and race: Verbal disputes and metaphysical deflationism." *Journal of Social Philosophy* 00 (2021): 1–19.

———. "Woman as a Politically Significant Term: A Solution to the Puzzle." *Hypatia* 31, no. 2 (2016): 245–258.

Doleac, Jennifer L., and Benjamin Hansen. "The Unintended Consequences of 'Ban the Box': Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden." *Journal of Labor Economics* 38, no. 2 (2020): 321–374.

Dretske, Fred. "Triggering and Structuring Causes." *A Companion to the Philosophy of Action* (2010): 139-144.

Eidelson, Benjamin. *Discrimination as Disrespect*. Oxford: Oxford University Press, 2015.

Equal Employment Opportunity Commission v. Sears, Roebuck & Co., 628 F. Supp. 1264 (N.D. Ill. 1986).

Equal Employment Opportunity Commission v. Sears, Roebuck & Co. 839 F.2d 302 (7th Cir. 1988).

Eronen, Markus. "Causal discovery and the problem of psychological interventions." *New Ideas in Psychology* 59, (2020): 100785.

Flake, Dallan F. "Do Ban-the-Box Laws Really Work?." *Iowa Law Review* 104, (2019): 1079–1127.

Franklin-Hall, L. R. "High-Level Explanation and the Interventionist's 'Variables Problem'." *British Journal of Philosophy of Science* 67, no. 2 (2016): 553–577.

Gaddis, Michael S. "An Introduction to Audit Studies in the Social Sciences." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. M. Gaddis. Springer, 2017.

Gheaus, Anca. "Gender Justice." *Journal of Ethics & Social Philosophy* 6, no. 1 (2011): 1–24.

Glennan, Stuart. "Mechanisms, causes, and the layered model of the world." *Philosophy and Phenomenological Research* 81, no. 2 (2010): 362-381.

Guryan, Jonathan, and Kerwin Kofi Charles. "Taste-based or statistical discrimination: The economics of discrimination returns to its roots." *The Economic Journal* 123, no. 572 (2013): F417–F432.

Haavelmo, Trygve. "The probability approach in econometrics." *Econometrica* 12, Supplement, iii–vi and 1–115.

———. "The statistical implications of a system of simultaneous equations." *Econometrica* 11, no. 1 (1943): 1–12.

Hall, Ned. "Causation and the Aims of Inquiry." In *Statistics and Causality: Methods for Applied Empirical research*, edited by Alexander von Eye and Wolfgang Wiedermann. Wiley 2016: 3–30.

———. "Structural Equations and Causation," *Philosophical Studies* 132, no. 1 (2007): 109–136.

———. "Structural Equations and Causation," (2006), manuscript.

Halpern, Joseph Y. "Appropriate Causal Models and the Stability of Causation." *The Review of Symbolic Logic* 9, no. 1 (2016): 76–102.

Halpern, Joseph Y., and Christopher Hitchcock. "Actual Causation and the Art of Modeling." In *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*. London: College Publications, 2010, 383–406.

———. "Causes and Explanations: A Structural-Model Approach. Part I: Causes." *The British Journal of Philosophy of Science* 56, no. 4 (2005): 843–887.

Haslanger, Sally. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press, 2012.

———. "What is a (social) structural explanation?." *Philosophical Studies* 173, no. 1 (2016): 113–130.

Heckman, James. "Detecting Discrimination." *Journal of Economic Perspectives* 12, no. 2 (1998): 101–116

Heckman, James, and Peter Siegelman. "The Urban Institute Audit Studies: Their Methods and Findings." In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael Fix and Raymond J. Struyk. Washington, D.C.: The Urban Institute Press, 1993, 187–258.

Hellman, Deborah. *When is Discrimination Wrong?*. Cambridge: Harvard University Press, 2008.

Hiddleston, Eric. "Review of Woodward, *Making Things Happen*." *The Philosophical Review* 114, no. 4 (2005): 545–547.

Hitchcock, Christopher. "Events and Times: A Case Study in Means-Ends Metaphysics." *Philosophical Studies* 160, no. 1 (2012): 79–96.

———. "Prevention, Preemption, and the Principle of Sufficient Reason." *The Philosophical Review* 116, no. 4 (2007): 495–532.

———. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy* 98, no. 6 (2001): 273–299.

———. "Three Concepts of Causation." *Philosophy Compass* 2, no. 3 (2007): 508–516, 510.

Holland, Paul W. "Causation and Race." *ETS Research Report Series* 2003, no. 1 (2003).

Jacquemet, Nicolas, and Constantine Yannelis. "Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market." *Labour Economics* 19, no. 6 (2012): 824–832.

Kment, Boris. "Causation: Determination and Difference-Making." *Noûs* 44, no. 1 (2010): 80-111.

Kohler-Hausmann, Issa. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking." *Northwestern Law Review* 113, no. 5 (2019): 1163–1227.

Lewis, David. "Causation." *The Journal of Philosophy* 70, no. 17 (1973): 556–567.

———. "Causation as Influence." *The Journal of Philosophy* 97, no. 4 (2000): 182–197.

MacDonald, Cynthia, and Graham MacDonald. "Mental Causes and Explanation of Action." *The Philosophical Quarterly* 36, no. 143 (1986): 145-158.

———. "The Metaphysics of Mental Causation." *The Journal of Philosophy* 103, no. 11 (2006): 539-576.

MacKinnon, Catherine A. "Difference and Dominance: On Sex Discrimination." In *Feminism Unmodified: Discourses on Life and Law,* Cambridge: Harvard University Press, 1984, 32–45.

Maudlin, Tim. "A modest proposal concerning laws, counterfactuals, and explanations." In *The Metaphysics Within Physics*, edited by Tim Maudlin. Oxford: Oxford University Press, 2007, 5–49.

Milkman, Ruth. "Women's History and the Sears Case." *Feminist Studies* 12, no. 2 (1986): 375–400.

Mill, John Stuart. *A System of Logic, Ratiocinative and Inductive. Vol. I*, Book III., Chapter VIII. London/New York: Longmans, Green, and Co., 1843.

Mills, Charles. "'Ideal Theory' as Ideology." *Hypatia* 20, no. 3 (2005): 165–184.

Moreau, Sophia. *Faces of Inequality: A Theory of Wrongful Discrimination.* Oxford: Oxford University Press, 2020.

Morrison, Toni. "A Humanist View." Portland State University, 1975. https://www.mackenzian.com/wp-content/uploads/2014/07/Transcript_Portland State_TMorrison.pdf.

Nagel, Thomas. "Equal Treatment and Compensatory Discrimination." *Philosophy & Public Affairs* 2, no. 4 (1973): 348–363

Nunley, John M., Adam Pugh, Nicholas Romero, and R. Alan Seals. "Unemployment, Underemployment, and Employment Opportunities: Results from a Correspondence Audit of the Labor Market for College Graduates." *ILR Review* 70, no. 3 (2017): 642–669.

Paul, L. A. "Aspect Causation." *The Journal of Philosophy* 97, no. 4 (2000): 235–256.

Paul, L. A., and Ned Hall. *Causation: A User's Guide*. Oxford University Press, 2013.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 2009 [2000].

Perry, Mark J. "Details in BLS Report Suggest That the 'Gender Earnings Gap' Can Be Explained by Age, Marital Status, Children, Hours Worked, Etc." American Enterprise Institute, March 6, 2021. https://www.aei.org/carpe-diem/details-in-bls-report-suggest-that-most-of-the-gender-earnings-gap-is-explained-by-age-marital-status-children-hours-worked/.

Prescott-Couch, Alexander. "Explanation and Manipulation." *Noûs* 51, no. 3 (2017): 585–520.

Rachels, James. "What People Deserve." In *Justice and Economic Distribution*, edited by John Arthur and William Shaw. Englewood Cliffs, New Jersey: Prentice-Hall, 1978, 150–163.

Rich, Stephen M. "Against Prejudice." *George Washington Law Review* 80, no. 1 (2011).

Schouten, Gina. *Liberalism, Neutrality, and the Gendered Division of Labor*. Oxford: Oxford University Press, 2019.

Sider, Ted. "Substantivity in Feminist Metaphysics." *Philosophical Studies* 174, no.1 (2017): 2467–78, 2473.

———. *Writing the Book of the World*. Oxford: Oxford University Press, 2011.

Simonsohn, Uri. "[51] Greg vs. Jamal: Why Didn't Bertrand and Mullainathan (2004) Replicate?" Data Colada, February 15, 2020. http://datacolada.org/51.

Spirtes, Peter, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Cambridge: MIT press, 2000 [1993].

Strevens, Michael. "Review of *Making Things Happen." Philosophy and Phenomenological Research* 74, no. 1 (2007): 233–249.

Thomson, Judith Jarvis. "Preferential Hiring." *Philosophy & Public Affairs* 2, no. 4 (1973): 364–384.

Turk, Katherine. *Equality on Trial.* Philadelphia: University of Pennsylvania Press, 2016.

Turner, Margery Austin, Michael Fix, and Raymond J. Struyk. *Opportunities denied, opportunities diminished: Racial discrimination in hiring.* Washington, D.C.: The Urban Institute Press, 1991.

Urquidez, Alberto G. "What accounts of 'racism' do." *The Journal of Value Inquiry* 52, no. 4 (2018): 437–455.

Weslake, Brad. "Exclusion Excluded." *International Journal for the Philosophy of Science*, forthcoming.

Williams, Patricia J. "Alchemical notes: Reconstructing ideals from deconstructed rights." *Harvard Civil Rights-Civil Liberties Law Review* 22, (1987): 401.

Wilson, Robert A. "Individualism, causal powers, and explanation." *Philosophical Studies* 68, no. 2 (1992): 103-139.

Woodward, James. "A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment)." *Philosophy of Science* 81, no. 5 (2014): 691–713.

———. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91, no. 2 (2015): 303–347.

———. *Making Things Happen.* Oxford: Oxford University Press, 2003.

———. "Mental Causation and Neural Mechanisms." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, edited by Jakob Hohwy and Jesper Kallestrup. Oxford: Oxford University Press, 2008.

———. "Methodology, Ontology, and Interventionism." *Synthese* 192, no. 11 (2014): 3577–99.

———. "The Problem of Variable Choice." *Synthese* 193, no. 4 (2016): 1047–1072.

Woodward, James, and Christopher Hitchcock. "Explanatory Generalizations, Part I: A Counterfactual Account." *Noûs* 37, no. 1 (2003): 1–24.