DOI: 10.1017/psa.2022.88

This is a manuscript accepted for publication in *Philosophy of Science*.

This version may be subject to change during the production process.

Well-Defined Interventions and Causal Variable Choice

Zili Dong

Department of Philosophy, Western University

1151 Richmond St, London, ON Canada, N6A 5B8

Email: zdong67@uwo.ca

Abstract

There has been much debate among scientists and philosophers about what it means for (hypothetical) interventions invoked in causal inference to be "well-defined" and how considerations of this sort should constrain the choice of causal variables. In this paper, I propose that an intervention is well-defined just in case the effect of interest is well-defined, and that the intervention can serve as a suitable means to identify that effect. Based on this proposal, I identify several types of ambiguous intervention. Implications for variable choice are discussed using case studies drawn from the sciences.

Acknowledgments

I would like to thank Wayne Myrvold for his valuable feedback. I am also grateful to James Woodward, two anonymous referees, Shimin Zhao, and participants at the Rotman Graduate Student Conference (RGSC2021), for their helpful comments and questions.

1. Introduction

Causal inference in science starts with selecting a set of relevant variables (**V**={*X*, *Y*, *Z*, …}). These variables are used to formulate our causal questions (e.g., "what is the effect of *X* on *Y*?"). If necessary, they can also be used to represent background causal knowledge (i.e., prior knowledge about causal relations among variables in **V**) or to provide the context of a causal study (by specifying important background factors). However, not all variables are suitable for answering causal questions. A poor choice of variables may lead to erroneous causal inference.

Consider an illustrative example. Suppose we want to investigate the effect of total cholesterol (*TC*) on heart disease (*HD*) by conducting a clinical trial. Through dietary control, patients in the treatment group maintain a high level of *TC*, and patients in the control group maintain a medium level. Results show that patients with medium *TC* are 20% less likely to develop heart disease. Can we give a *causal* interpretation of this association?

No. Total cholesterol has two major components: low-density lipoprotein (*LDL*) and high-density lipoprotein (*HDL*). A medium level of *TC* can be realized in multiple ways (e.g., low *HDL* + high *LDL* or medium *HDL* + medium *LDL*). Moreover, *HDL* protects us against heart disease, whereas *LDL* does the opposite. Consequently, different realizations of the same level of *TC* may have significantly different effects on *HD*. For this reason, the above-

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

mentioned manipulation of *TC* is *ambiguous* (Spirtes & Scheines 2004). Note that the ambiguity will remain even if we assume there is no confounding or random error. The source of the problem is that *TC* is unsuitable for being a cause in the above study.

The above example invites a general question: when is it appropriate to use a variable (relative to a variable set) as a cause in causal inference? I will confine my discussion to a broadly construed *interventionist* approach to causation. Interventions, whether experimental or hypothetical, are useful in identifying causal effects. However, for some variables, interventions on them may provide misleading causal information. This motivates practitioners and philosophers of causal inference (e.g., Holland 1986, 2008; Hernán & Taubman 2008; VanderWeele 2018; Woodward 2016) to adopt the following constraint on variable choice: A variable X is suitable for investigating its effect on Y only if we have *well-defined interventions* (WDIs) on X with respect to Y.[1]

There has been much debate on what kind of interventions should count as WDIs and how exactly considerations of this sort could constrain the choice of causal variables. A view that has been influential among some practitioners is that only treatments can be causes since only treatments can support WDIs (e.g., Holland 1986, 2008). However, more and more researchers have contended that the scope of WDIs should be broadened, and accordingly, we

[1] This condition is nevertheless insufficient since there are other considerations in variable choice. See Woodward (2016).

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

should be more liberal on causal variable choice (see, e.g., Glymour & Glymour 2014; Glymour & Spiegelman 2017; Marcellesi 2013; Pearl 2018; Schwartz et al. 2016).

It would be most desirable if we could find a rigorous definition of WDIs and an expedient recipe for selecting cause variables so as to settle the debate once and for all. I doubt this could ever be achieved, for reasons we shall see soon. That said, this does not mean we cannot say anything about WDIs that is of general importance. My strategy in this paper is to first formulate a general, though sketchy, characterization of WDIs and then flesh it out by analyzing a few specific cases in which interventions are, or appear to be, ill-defined.

The paper is structured as follows. Section 2 introduces a definition of causal effects using Woodward's (2003) notion of ideal intervention. In section 3, I propose that an intervention on X with respect to Y is well-defined, if and only if 1) the effect of interest is well-defined (under conceivable ideal interventions on X), and 2) all things considered, the intervention can be used to reliably identify that effect (either directly or through some indirect methods). Based on this proposal, in section 4, I show interventions may (appear to) be ill-defined or ambiguous for various reasons. An intervention may be ill-defined if the effect of interest is ill-defined. Alternatively, an intervention may be ill-defined because it is considered unsuitable for identifying an effect of interest. In particular, I emphasize that different types of ambiguous interventions require different ways of handling them and have different implications for variable choice. To illustrate these ideas, I examine a few typical examples of ambiguous interventions (e.g., interventions on total cholesterol, obesity, and race).

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

It is worth emphasizing that causal variable choice is a highly pragmatic issue; a set of variables may be suitable for being causal variables in one study or context but not in another. This is why I believe, instead of asking "What can be a cause?" or "Is *X* a cause?" a better question to ask is, "Should we use *X* as a cause in a certain (type of) causal study?" The latter question makes it clear that whether *X* should be used as a cause is contingent on the context and aim of a study, the expected advantages and disadvantages of using *X*, and so forth. For this reason, the primary aim of this paper is not to reach definite conclusions on which variables can be causes but to illustrate the nuances and complexities of causal variable choice.

2. Causation and Intervention

There is a close link between the notion of causation and that of intervention, which enables us to define and identify causation through interventions. A broadly construed interventionist approach to causation is widely endorsed or presupposed by philosophers and practitioners of causal inference. In particular, it can be seen as a conceptual foundation for two major causal inference frameworks used in the medical and social sciences: the causal-modeling framework (Pearl 2009; Spirtes et al. 2000) and the potential-outcomes framework

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

(Rubin 1974; Holland 1986). Issues discussed in my paper arise for both, but I will set up the discussion using the causal-modeling framework.[2]

The primary goal of causal inference in the medical and social sciences is to identify the size of an effect of interest. The *effect* of X on Y or CE(X, Y) (by default, this means the total effect) is defined in terms of the difference in Y's value or probability distribution when X is set to different values by ideal interventions.

Let me unpack this definition. First, it works only when the interventions in question are ideal. An intervention I on X with respect to Y is an *ideal intervention* if and only if (Woodward 2003, 98):

I. I causes X such that it renders X independent of its other causes (i.e., I breaks all the other arrows going into X).

[2] Here are some reasons why I prefer the causal-modeling framework. As we shall see, the definition of effects in terms of ideal interventions from the causal-modeling framework will prove essential in understanding WDIs. This definition does not have an exact counterpart in the potential-outcomes framework. The latter framework originated from an extension of causal reasoning in randomized trials to observational studies (Rubin 1974), and it has mainly been used for purposes of policymaking. As a consequence, the potential-outcomes community tends to restrict the scope of causes to treatments (more on this below).

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

II. *I* does not cause *Y* through a causal path that does not go through *X*.[3]

III. *I* is (statistically) independent of any variable *Z* that causes *Y* and that is on a directed path that does not go through *X* (assuming the common cause principle, this means *I* and *Y* share no common causes).

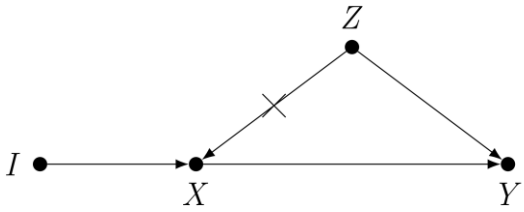

Figure 1. An ideal intervention *I* on *X* with respect to *Y* breaks the arrow going into *X*.

When there are ideal interventions on *X* with respect to *Y*, the association between *X* and *Y* observed under these interventions can be straightforwardly interpreted as the effect of *X* on *Y*. Hence, it is most natural and convenient to *define* causal effects in terms of ideal

[3] This definition assumes that variables used in a causal model do not stand in *noncausal* (e.g., logical or ontological) relationships. But sometimes we do want to include noncausal relationships in our causal models; e.g., we may add one of *X*'s supervenience bases, *X**, to the causal graph in figure 1, to show how $X \rightarrow Y$ is realized by lower-level causation. In this new causal graph, condition (II) needs to be revised such that the additional path "$I \rightarrow X^* \rightarrow Y$" does *not* count as a violation of condition (II) (see Woodward 2015). I add this note because my later discussion will make use of causal graphs that contain noncausal relationships.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

interventions. Note, for definitional purposes, these ideal interventions need only be conceivable (we will get clearer on this later). They can be merely *hypothetical*—they do not have to be implemented in practice, nor do they have to be feasible for human agents.

Moreover, the above definition only talks about total effects, but sometimes we are not (just) interested in total effects. It is often the case that we are interested in causal models in which X causes Y through several paths. The total effect of X on Y can then be decomposed along these paths. What is of particular interest is the decomposition of a total effect into a direct and an indirect effect. The direct effect of X on Y is exerted through a direct causal path from X to Y without intermediate variables. It is defined as the effect of intervening on X with respect to Y when holding all the other variables in the model (besides X and Y) fixed by interventions.

Lastly, in the medical and social sciences, we typically need to estimate average causal effects from population-level data. For example, in a clinical trial aimed at identifying CE(X, Y), we can randomly assign participants to the treatment group (X=1) or the control group (X=0); as a result, the other causes of Y besides X will be distributed similarly across the two groups. We can then infer CE(X, Y) by seeing how the difference in X's value leads to a difference in Y's probability distribution between the two groups. In addition to experiments, we may also identify effects from observational data. Here we need to consider hypothetical interventions: CE(X, Y) is identifiable, if we can infer from observational data and our background causal knowledge what the association between X and Y would be, had there been

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

ideal interventions on X. It can be shown that with enough background knowledge, CE(X, Y) can be identified from observations by adjusting for confounding (see Pearl 2009).

3. Characterizing Well-Defined Interventions

In section 3.1, I propose a general characterization of WDIs. In section 3.2, I compare my proposal with existing ones in the literature.

3.1 WDIs as Interventions Suitable for Identifying Well-Defined Effects

We have introduced the notion of ideal interventions to define causal effects. What does this have to do with WDIs? It may be tempting to think that an intervention is well-defined just in case it is an ideal intervention. However, despite both being regulative ideals, the two notions of intervention should be distinguished. As shown above, the notion of ideal intervention is the key to a general graphical definition of effects in the causal-modeling framework. In this framework, causal effects and ideal interventions are both defined with respect to a set of chosen variables that are *assumed* to be suitable for causal inquiry. This assumption is important since if the variables are poorly chosen (e.g., when X and Y are logically related), these definitions will no longer be applicable, and CE(X, Y) will be undefined. Nevertheless, the standard causal-modeling framework does not come with a procedure to verify the suitableness of a variable choice.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

In causal inference practice, where the goal is to identify a particular effect of interest, that a variable choice is suitable for the intended goal is something practitioners need to verify rather than assume. The idea of WDIs is concerned precisely with this issue. If we want to investigate the effect of X on Y—either by experimentally manipulating X or by considering a hypothetical intervention implied in observations—we should ensure that the intervention we use to identify CE(X, Y) is in some sense "well-suited" for the job. This is a complicated issue that involves the specific nature of the variables in question, the researcher's goals or interests, the resources and data available to us, and so on. Therefore, it is unlikely that we can give a general and rigorous "definition" of WDIs.

That being said, the notion of ideal interventions and that of WDIs are closely related. The reason we want WDIs in causal inference is that we want to identify an effect of interest reliably. Naturally, this requires that, first of all, the effect of interest be well-defined—under ideal interventions. After confirming that the effect of interest is well-defined, we also need to ensure that the intervention we have at hand is suitable for identifying that effect. In the best scenario, if the available intervention is ideal, we can readily identify the effect of interest. But ideal interventions are not always available, which complicates things. Luckily, non-ideal interventions can also be well-suited for identifying effects; sometimes, they are even more desirable when practical concerns are considered. In some cases, data about ideal interventions are not available, so non-ideal interventions will be the only option left. I will come back to these issues in section 4.1.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Given what has been said, we are in a position to formulate the following characterization of WDIs:

> **Well-Defined Interventions:** An (hypothetical) intervention on X, with CE(X, Y) as the effect of interest, is well-defined for a certain (type of) causal study if and only if 1) CE(X, Y) is (sufficiently) well-defined or determinate under conceivable ideal interventions on X with respect to Y; and 2) given the resources we have, we can identify CE(X, Y) from the intervention in a reasonably unambiguous way, either directly or through some indirect method.

In short, WDIs are *interventions that are good for identifying well-defined effects*. Some remarks are in order. First, although this sketchy characterization is a good start, it is nothing like a rigorous "definition" of WDIs. It leaves much to be said. Perhaps what is more important for our discussion are those more concrete situations in which an intervention fails to satisfy the above conditions. Regarding an intervention suspected to be ill-defined or ambiguous, we may ask: what is the source of the ambiguity? Can the ambiguity be avoided or reduced to some reasonable degree? What does it tell us about variable choice? As we shall see in section 4, interventions may be ill-defined for various reasons. There, I will examine some typical examples of ambiguous interventions and show, in each case, how the ambiguity may be dealt with.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Another thing to note is that although statistical evidence about (hypothetical) interventions is probably the most important source of evidence for causal inference (especially in the medical and social sciences), they are not the only source. Considerations about theories, laws, mechanisms, and the like can also provide valuable information about causal effects.[4] In particular, such knowledge is essential for diagnosing whether CE(X, Y) is well-defined under conceivable ideal interventions.

If we have rigorous theories or laws featuring X and Y, especially in the physical sciences, CE(X, Y) can be inferred from such knowledge since, in this case, ideal interventions on X are relatively easy to conceive (even if they are physically impossible). Consider, for example, the moon's gravitational effect on the earth's tides if the distance between the moon and the earth were doubled. Woodward (2003, 129ff; 2008) argues that even if all physically possible interventions on the distance are non-ideal, we can still conceive an ideal intervention on the distance, given Newtonian mechanics. Under this ideal intervention, the moon's gravitational effect on the earth's tides is well-defined.

In special sciences, however, rigorous theories or laws are rare. What we are likely to have is knowledge about causal mechanisms underlying causal relationships. For example, if X affects Y through some underlying physiological mechanisms, knowledge about these

[4] Regardless of whether they can be subsumed under a broadly construed interventionist framework, the point here is that these considerations play different and complementary roles in causal inference that cannot be replaced by statistical evidence about interventions.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

mechanisms may suggest at least qualitative information about CE(X, Y). Such information may not directly tell us what CE(X, Y) is, but it can be essential in diagnosing whether CE(X, Y) is well-defined, especially in cases where humanly feasible interventions are always non-ideal. In such cases, knowledge about the underlying mechanisms can help us conceive ideal interventions on X so as to determine whether CE(X, Y) is sufficiently well-defined under these interventions.

Finally, I want to highlight a sometimes-neglected point: an intervention must be stated relative to *an effect of interest*: whether an intervention is well-defined depends on which effect we are interested in. When CE(X, Y) is well-defined whereas CE(X, Z) is not, interventions on X may be well-defined with respect to Y but ill-defined with respect to Z. For this reason, the fact that interventions on X are well-defined in some cases cannot directly justify using X as a cause in general. Similarly, the fact that interventions on X are ill-defined in one case may not be a reason to abandon the use of X as a cause altogether. For example, even if manipulations of *TC* with respect to *HD* are ambiguous, this does not mean manipulations of *TC* with respect to other health outcomes are ill-defined as well.

In some cases, it may be that interventions on X with respect to Y are ill-defined when the effect of interest is the direct effect of X on Y, whereas they are well-defined when the effect of interest is the total effect. In such types of cases, it is important to be clear about whether the effect of interest is the direct or the total effect. For a concrete example, consider sex. For the sake of argument, let us suppose that both the direct and total effects of sex on

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

one's choice of college major are well-defined. Still, whether hypothetical interventions on sex implied in observations are well-defined depends on one's effect of interest. These interventions may be well-defined if we are interested in the total effect. But one may also be interested in the direct effect of sex, with intermediate variables like sex biases held fixed. In that case, however, I suspect these interventions are no longer well-defined since they are not good for identifying the effect of interest. This is because, as far as I can see, it is unlikely that we can fully adjust for intermediate variables, given that all the data we have access to come from a society where sex biases constantly exist.

3.2 Comparisons with Other Proposals

Holland's proposal. Holland distinguishes between treatments (e.g., drugs, diets, education) and attributes (e.g., academic performance, obesity, race). As I understand it, a treatment is a substance or an operation that can be defined and manipulated independently of the subject to be treated, whereas an attribute (e.g., being a woman or being Black) is constitutive of a subject, and it cannot be manipulated without simultaneously manipulating other characteristics of the subject. For this reason, Holland thinks only treatments are "manipulable" (more precisely, what he actually means is that only treatments can be manipulated unambiguously).

Therefore, for Holland, interventions on X are WDIs only if X can, in principle, be used as a *treatment* variable in a (randomized) controlled experiment. For this reason, "causes are

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

only those things that could, in principle, be treatments in experiments" (Holland 1986, 954; here "in principle" emphasizes that this is not a matter of practical feasibility). This proposal does have its reasonableness. In a well-conducted experiment, random assignments of treatments are designed to be ideal interventions. In particular, condition (II) is automatically satisfied, since there is no (non-trivial) additional causal path from a treatment assignment to the outcome of interest.

A major limitation of Holland's proposal, however, is that it is overly *restrictive*, since it rules out all attributes as causes. Contrary to Holland's proposal, it is common practice to take physical, biological, or psychological attributes as causes in the sciences. Many of these attributes (such as mass, chemical compositions, at least some physiological states, and perhaps some mental attributes) have well-defined effects. We also have good means to identify these effects, even when interventions on these attributes are typically not ideal. It is overly conservative to exclude all attributes from the realm of causes.

The consistency proposal. A more recent proposal says that a (hypothetical) intervention that sets X to x is well-defined with respect to Y if and only if the outcome of the setting is *consistent* among different realizations of the setting; that is, different versions of setting X to x must determine a (sufficiently) unique probability distribution of Y (see Hernán & Taubman 2008; Hernán 2016; VanderWeele 2018).

Due to considerations related to generalizability or transportability, consistency has been taken as a basic assumption or requirement in the potential-outcomes framework. At first

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

sight, this also seems to be a reasonable requirement. Suppose, according to two studies on the same population, setting X to x leads to different distributions of Y. It is not immediately clear what can be learned from these inconsistent results. However, the inconsistency we see here is merely a symptom that may have several possible sources. Two causal studies may report inconsistent results when CE(X, Y) is inconsistent (e.g., when X is a heterogeneous variable). Alternatively, assuming CE(X, Y) is consistent, two studies may still report apparently "inconsistent" results if at least one result comes from an intervention that is not ideal.

These two types of inconsistency should be clearly distinguished. In the first case, to avoid inconsistency, we often have to stop using X as a cause or replace it with other variables. In the latter case, the reported "inconsistency" does not indicate that CE(X, Y) is ill-defined, nor does it directly imply that X is unsuitable for being a cause. The apparent inconsistency can be resolved as long as we can still consistently identify CE(X, Y), through some indirect methods, from the non-ideal interventions used in the studies. That is, at least when the choice of cause variable is concerned, it is not the apparent (in)consistency between different causal studies that matters but the (in)consistency of the effect of interest. This point will be further explained in section 4.2.3 with a case study.

Pearl's proposal. Pearl (2018) emphasizes that we should define CE(X, Y) in terms of "an ideal, atomic intervention" on X, represented as $do(X)$. Specific or realistic interventions implemented in experiments or implied in observations tend to be non-atomic or "imperfect." For this reason, they may report seemingly inconsistent results. But this has nothing to do with

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

CE(X, Y), which is defined in terms of an atomic or surgical intervention on X, namely, $do(X)$. Pearl's $do(X)$ is roughly equivalent to Woodward's ideal intervention—the latter can be seen as an explication of $do(X)$. Given this common ground, I think Pearl would agree with me on what I have said about WDIs.

My main complaint about Pearl (2018) is that he seems to have underestimated the complexity of the issue. He advocates a liberal attitude towards the use of variables in causal inference, and particularly, he seems to assume race can be treated on a par with biological attributes (e.g., obesity). I will argue later that we should be more cautious here, especially regarding race. Interventions can appear ambiguous for various reasons. Sometimes the ambiguities are merely apparent, but sometimes they can pose genuine problems.

In the causal discovery literature, interventionist considerations have also been considered essential for selecting causal variables (see Chalupka, Eberhardt & Perona 2016, 2017). For example, Chalupka et al. (2017, 140) require that "causal variables should permit well-defined experimental interventions." Their project aims to develop a domain-general framework for constructing high-level causal variables from low-level ones so as to discover high-level causation from low-level data. The problem addressed in my paper differs from theirs in important ways. Here, we are concerned not with the construction of causal variables but with extant and already widely used high-level causal variables in the sciences whose causal status nevertheless remains controversial, such as obesity, race, and so on. To me, this

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

latter issue resists a domain-general solution, since variables in different domains may very well be afflicted by different types of ambiguous interventions.

4. Ill-Defined Interventions and Implications for Variable Choice

In this section, I will first discuss, in general, when an intervention in a causal study may be judged ill-defined or ambiguous. Then, I will study a few typical cases of ambiguous interventions and discuss the implications for variable choice.

4.1 When Is an Intervention Ill-Defined?

Let us first consider the situation in which the effect of interest CE(X, Y) is (sufficiently) *well-defined*. It would be best if the intervention used to identify CE(X, Y) is ideal; there is little doubt that such an intervention is a WDI. When the intervention is *not* ideal, the observed association between X and Y cannot be directly interpreted as CE(X, Y). Now, if one sticks to the old wisdom and thinks only ideal interventions such as randomized trials can provide reliable causal inference, they may categorize non-ideal interventions as ill-defined. In what follows, I am going to show this is wrong. Besides, in cases where CE(X, Y) is well-defined, I see no difficulty in saying that, by default, we can use $\{X, Y\}$ as causal variables. We are required to stop using them as causal variables when, for whatever reasons, the intervention available to us cannot be used to reliably identify CE(X, Y).

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

An intervention can fail to be ideal by violating any of the three conditions for an ideal intervention (see section 2). When an intervention violates two or even three of the conditions, I suspect that such a poor intervention is unlikely to be a WDI since, in practice, it would be difficult to identify the effect of interest from it. So, in what follows, I will discuss interventions that violate only one of the three conditions.

Among them, interventions violating condition (II) are of particular interest to us because these interventions have been explicitly categorized as ambiguous by practitioners of causal inference from the potential-outcomes community. This type of intervention will be discussed in detail in section 4.2.3.

Condition (I) for an ideal intervention says that an ideal intervention on X should break all the other arrows going into X. Those interventions that fail to break arrows going into X have been called "soft interventions" in the causal discovery literature (Eberhardt & Scheines 2007) or "(causal) instruments" in econometrics (Reiss 2005). It has been shown that, when used correctly, a soft intervention or a causal instrument can provide reliable information about causal structures or effects. If this is true, certainly non-arrow-breaking interventions are qualified as WDIs. In fact, when ideal interventions are unavailable or unethical, non-arrow-breaking interventions may be the only option.

Interventions that violate condition (III) are typically regarded as "ill-defined" in randomized experiments, since the violation suggests that the treatment assignments are not properly randomized as designed. These interventions are therefore confounded. But this does

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

not mean that these interventions can no longer be suitable means for identifying causal effects. For example, Pearl (2019, p. 4) shows that we may be able to use interventions that violate condition (III) as indirect tests for CE(X, Y) when it is otherwise unidentifiable. For this reason, we should not say interventions violating condition (III) are ill-defined simply because they are not the type of intervention we normally want in randomized experiments.

Now, consider the other situation in which CE(X, Y) is *ill-defined*. In this situation, all interventions on X with respect to Y are ill-defined. Prima facie, this indicates that X is not suitable for being a cause of Y, and we should be prepared to stop using $\{X, Y\}$ as causal variables—unless the ambiguity in CE(X, Y) is acceptable or can be somehow managed. There can be many ways CE(X, Y) is ill-defined; here, I consider two.

First, CE(X, Y) is ill-defined when X is a heterogeneous variable with respect to Y. In that case, even ideal interventions on X will be ambiguous (as we saw with *TC* and *HD*). Whether this means we should abandon X depends on the degree of ambiguity and whether the ambiguity can be reduced. Typically, heterogeneous variables are also variables that are easier to measure, more general, and so on. So, there is often a trade-off between the reduction of ambiguity and other desiderata. When the disadvantages outweigh the advantages of using X as a causal variable, we should stop using X or replace it with other variables.

Second, CE(X, Y) will be ill-defined when there are conceptual or ontological disputes over the concept we use to construct X such that there is no univocal causal structure underlying an ideal intervention on X with respect to Y. In other words, due to conceptual

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

ambiguities in *X*, an ideal intervention on *X* with respect to *Y* can be interpreted as manipulations of different sets of (lower-level) variables, depending on how *X* is understood. If this happens, it seems unwise to continue using *X* as a causal variable. One may attempt to resolve or circumvent the conceptual ambiguities by introducing a simplified or idealized version of *X*. However, as we shall see below, the trouble is that it might turn out that this "simplified-*X*" is no longer the same variable as the one we are initially interested in, which in effect justifies the need to replace *X* with some other variable(s).

4.2 Kinds of Ill-Defined Interventions: Case Studies

1) *Interventions on Heterogeneous Variables*

X is a heterogeneous variable with respect to *Y* if *X*'s obtaining value *x* can be realized in multiple ways such that different ways of setting *X* to *x* by interventions result in significantly different probability distributions of *Y*, even when the interventions are assumed to be ideal.[5] In this case, interventions on *X* with respect to *Y* will be ambiguous, since CE(*X*,

[5] Another possible scenario in which *X* does not have a determinate effect on *Y* is when the outcome variable *Y* is heterogeneous. For example, Hamer et al. (2021) criticize a recent study on the genetic causes of human homosexuality on the grounds that the study classifies individuals as "homosexuals" as long as they have *ever* engaged in same-sex sexual behavior. But this will create a highly heterogeneous group of "homosexuals" which may result in misleading associations between (alleged) genetic causes and "homosexuality."

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Y) is ill-defined. More specifically, a variable *X* can be heterogeneous with respect to *Y* at least in the following two ways: 1) *X* is an aggregation of several variables with differential effects on *Y*; 2) the effect of *X* on *Y* differs in different units or contexts.[6]

In practice, we do not always know beforehand whether *X* is heterogeneous. We might first find out that interventions on *X* are ambiguous and then diagnose the ambiguity as a consequence of using heterogeneous variables. For example, imagine that back in the 1950s, we did not know the components of total cholesterol, and different studies reported inconsistent results about the effect of *TC* on *HD*. One plausible explanation for this inconsistency would be that *TC* was a heterogeneous variable.

Both types of heterogeneity mentioned above imply that high-level or population-level variables are more likely to be heterogeneous. But this does not mean we should always choose variables that are as fine-grained as possible. Coarse-grained variables have their advantages. For example, they reveal causal patterns that cover a broad range of phenomena, they can simplify causal models, they are easier to measure, and so on. Therefore, in practice, there is usually a *trade-off* between the reduction of heterogeneity and other goals in variable choice. When the heterogeneity is within reasonable limits, we need not have to abandon the variable.

---

[6] Although philosophers are more interested in the first type of heterogeneity, note that when scientists talk about "heterogeneous (treatment) effects," they are referring to the second type.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Suppose the heterogeneity in an aggregated variable X (= $X_1 + X_2$) with respect to Y does exceed a reasonable degree. In this case, the association between X and Y no longer supports causal interpretation, and we should use $\{X_1, X_2, Y\}$ as our new causal variables. Note that the (spurious) association between X and Y is not a consequence of confounding and cannot be eliminated by blindly applying confounding adjustment—not all systematic spurious associations are results of confounding. Sometimes they are brought about by poor variable choice and can only be eliminated by re-selecting our variables.

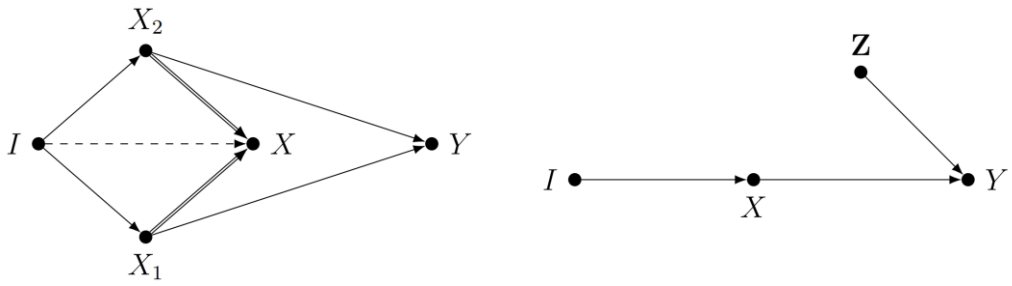

Figure 2. Two kinds of heterogeneity in CE(X, Y). (a) X is an aggregation of X_1 and X_2 with different effects on Y. (b) Background factors **Z** (which may differ in different units) are responsible for the heterogeneity.

To make the above point clearer, in figure 2(a), I draw a causal graph in which X and $\{X_1, X_2\}$ are included in the same causal model ("⇒" represents noncausal relations). In this graph, it makes no sense to disambiguate the intervention on X by holding X_1 fixed since the intervention will then collapse into an intervention on X_2. That is, holding X_1 fixed is in effect replacing X with $\{X_1, X_2\}$.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Another way X can be heterogeneous is when X's obtaining the same value can affect Y differently in different units or contexts. The situation can be represented schematically in figure 2(b).[7] In this graph, **Z** is a set of factors responsible for the heterogeneity in CE(X, Y). This kind of heterogeneity is related to the well-known fact that how X affects Y often depends on or interacts with the context or background factors (e.g., whether a drug can relieve my headache depends on various characteristics of mine).[8] Probably all high-level causal variables are more or less heterogeneous in this sense. It is virtually impossible for causal effects in the high-level sciences to remain consistent in all contexts. Luckily, the problem is not as serious as it appears.

In many cases, CE(X, Y) is still reasonably homogeneous in the "normal" or "average" causal contexts. X and Y can be perfectly legitimate causal variables if they are used in these contexts. Moreover, we may also be able to reduce the heterogeneity by holding variables in **Z**

[7] Note that although causal graphs like this one are instructive, they tend to be overly simplified. For example, as a special case, when X positively affects Y in half of the target population but negatively affects Y in the other half, it may turn out that CE(X, Y)=0 at the population level. In this case, it may be misleading to not draw an arrow from X to Y (and is probably equally misleading if we draw one). This special case, however, does not suggest that causal graphs are not useful, but it does show that it is important to attend to the heterogeneity in CE(X, Y). I thank an anonymous referee for pointing out this issue.

[8] I thank an anonymous referee for raising this point.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

constant (e.g., by studying units with similar levels of **Z**). The moral is that although the effects of higher-level variables tend to be heterogeneous due to context-dependency, many of them still describe robust properties of the world and can be useful in causal inference.

Consider biodiversity. According to McCann (2000), observations show that diversity tends to be positively correlated with ecosystem stability in an ecosystem. However, as McCann emphasizes, this association does not mean that diversity is the "driver" of ecosystem stability; that is, this association does not support a causal interpretation. This is because biodiversity is an abstract characteristic of an ecosystem that does not consider other factors that may contribute to stability, such as underlying interactions between species. Two habitats may have the same level of biodiversity but differ significantly in their modes of species interactions and hence differ in their levels of stability. As a result, attempts to manipulate ecosystem stability by manipulating biodiversity may fail if we ignore other relevant factors. Nevertheless, despite the above problem, biodiversity may still be seen as a well-defined cause of stability if we focus on ecosystems with roughly similar modes of species interactions. We may also successfully manipulate ecosystem stability by manipulating biodiversity if species interactions and other relevant factors are maintained at a fixed level.

2) *Interventions on Ontologically Controversial Variables*

Consider the following situation: CE(X, Y) is ill-defined because we cannot tell a *univocal* underlying story about how ideal interventions on X with respect to Y work. For

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

example, suppose, according to one plausible understanding, X supervenes on (or is defined by) variables $\mathbf{X}=\{A, B, C\}$, and therefore, an intervention on X is just an intervention on variables in $\mathbf{X}$. But according to another plausible understanding, X supervenes on $\mathbf{X^*}=\{A, B\}$. In this case, an intervention on X is just an intervention on $\mathbf{X^*}$. It follows that an ideal intervention on X does not have a determinate underlying causal mechanism. Now, if C has a significant effect on Y, it follows that CE(X, Y) is ill-defined. Any intervention on X with respect to Y will then be ill-defined.

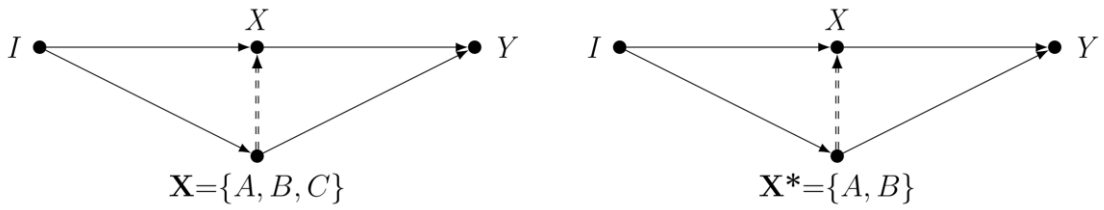

Figure 3. Causal graphs in which X supervenes on a set of lower-level variables. (a) The supervenience base is $\mathbf{X}=\{A, B, C\}$; (b) The supervenience base is $\mathbf{X^*}=\{A, B\}$.

This kind of ambiguity occurs when the concept used to construct X is conceptually contested, and this problem is especially severe in the social sciences. It is widely recognized that social categories are prone to conceptual disputes: they often have unclear boundaries, and their boundaries may even change over time (see Greene 2020 on how this may have an impact on the choice of causal variables in the social sciences); moreover, classifications of social groups themselves can affect how people behave (the so-called "looping effects"). In particular, for many social characteristics, there can be different conceptions of them that are

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

all, to some extent, plausible, and it is unlikely that philosophers and social scientists can achieve agreement on what these characteristics "really" are. This raises the question of whether ideal interventions on them are sufficiently univocal. Importantly, the issue here is not that we lack knowledge about these characteristics (although this is not to say empirical evidence is irrelevant to this issue). The controversies over them are primarily ontological rather than epistemic.

Take "race" as an example.[9] It is still heatedly debated what race is—whether race is biologically grounded, socially constructed, or a bit of both (James & Burgos 2020). An ideal intervention on race may mean different things based on how one conceives of race. Here, I will not take a stance on how we should understand race. For my discussion below, it suffices to acknowledge that so far, we do not (and probably never will) have a consensus on the ontological status of race.

Let us start with the suggestion that race can be biologically defined, and thus we can manipulate one's race through genetic engineering. We should immediately notice the biological indeterminacy involved in such a manipulation. Unlike sex, which is normally determined by the sex chromosomes, genetic variations between different human races and within each human race are far more complicated. There is probably no shared set of alleles that can unequivocally identify a particular race. Worse, the genetic variation within each race

[9] Feedback from Wayne Myrvold, James Woodward, and participants at RGSC2021 has significantly improved my discussion on race.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

may be even larger than the average variation between races. Therefore, it is difficult to determine which genes to modify when conceiving an ideal intervention on race. This will lead to ambiguities in interventions on race, since manipulating different sets of genes may have quite different effects on the outcome of interest.

Partly due to the above considerations, many people contend that race is not a well-defined biological category. If they are right, merely manipulating genetic factors is not a manipulation of race per se, but, at best, a manipulation of typical biological characteristics associated with race. An intervention on race needs more than this. If belonging to a racial group means receiving a particular kind of cultural upbringing, a manipulation of a Black girl's race should also change her cultural upbringing as well. For example, Marcellesi (2013) suggests that to manipulate race, besides genetic factors, we should also change environmental factors (e.g., who will be the mother of an embryo).

Another controversy concerns whether race is, by definition, genealogical. If race is not genealogical, a manipulation of a Black girl's race has nothing to do with her ancestors' race. In this case, given that her ancestors' race is a cause of her race, the manipulation will simply break the arrow from her ancestors' race to her race. Alternatively, one may contend that race is genealogical (e.g., one may say, for a person to be Black, at least one of their parents should be Black). In that case, an intervention on the girl's race will simultaneously be an intervention on her ancestors' race, since what we have here is a definitional relationship

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

between these variables. Consequently, depending on whether one thinks race is genealogical, an ideal intervention on race can have quite different causal structures.

I believe the above discussion suffices to show that, by default, we should avoid using race as a cause. Those who do not want to accept this conclusion are obliged to provide a way to disambiguate the concept of race so as to reduce the ambiguity in interventions on race. A straightforward way to do so is to consider a well-controlled, idealized scenario. Consider the following argument from Marcellesi (2013). Suppose in an imaginary society, there are two racial groups, A and B. All the individuals in the population are perfectly homogeneous regarding possible causes of wages other than race (e.g., education and working experience). Now suppose A-group is much less likely to get high-wage jobs than B-group. The only cause of this wage gap, Marcellesi argues, is race.

Marcellesi's argument relies on constructing a highly idealized scenario where conceptual or ontological disputes over race no longer exist. In this scenario, whether a person belongs to a racial group is a standalone fact that does not depend on other properties of this person since everyone is otherwise similar. However, even if we grant that in this imaginary scenario, race is a cause, the argument's relevance to a *realistic* society is unclear. In our society, different racial groups are heterogeneous in various properties, which is precisely why it is so hard to define race. The realistic concept of race is thus quite different from the one in Marcellesi's imaginary society. The upshot is that the concept of race may be inherently ambiguous so that if we try to make race a well-defined cause by idealizing its context of use,

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

it could turn out that this simplified concept of "race" is no longer the one we are really interested in.

Holland (2008) also argues against using race as a causal variable, but I disagree with him on the rationale behind his assessment: he thinks race is not a cause, since, as an attribute, race cannot be manipulated in the way we manipulate a treatment. Indeed, by restricting the realm of causes to treatments, the kind of ambiguity we find in social attributes like race will disappear, since there are rarely conceptual or ontological disputes over the nature of a treatment (which is typically defined operationally). However, as I have argued earlier, the limitation of this strategy is also severe: it not only rules out race as a cause but also any variable that describes an attribute, regardless of whether such an attribute is ontologically controversial.

In response to Holland (2008), Marcellesi (2013) contends that race can be used as a treatment: we can randomly assign race to embryos since we can assign biological factors (via genetic engineering) and environmental factors (by swapping embryos between mothers) to embryos. I suspect that this is a misunderstanding of the notion of "treatment" (see my discussion in section 3.2). Here, it seems what is being assigned as a treatment to an embryo is not race itself, but the procedure of genetic engineering and the embryo's upbringing.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

3) *Interventions that Are Ham-Handed (AKA Fat-Handed)*

Consider the intervention *I* that is shown in figure 4. It affects *Y* through an additional causal path that does not go through *X*. The association we observe between *X* and *Y* under this intervention is not identical to CE(*X*, *Y*). This intervention is hence *ham-handed* (AKA fat-handed).

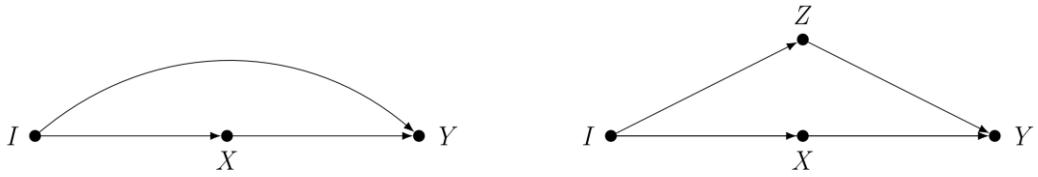

Figure 4. *I* is a ham-handed intervention on *X*. (a) *I* is a direct cause of *Y*. (b) *Z* is an intermediate variable between *I* and *Y*.

Hernán and Taubman (2008) state that ham-handed interventions are not WDIs. When interventions are ham-handed, two different methods of manipulating *X* may lead to different distributions of *Y* if, in the two methods, the effect of *I* on *Y* via the ham-handed causal path (or the "side effect" of *I*) differs significantly. Consider interventions on obesity. We can conduct two trials in which we manipulate the level of obesity from high to low but in different ways—through physical exercise and diet control, respectively. According to Hernán and Taubman, these two manipulations of obesity will lead to significantly different mortality outcomes because the side effects differ in the two studies; the mortality rate will be lower in the first trial. To them, this means that interventions on obesity violate the consistency

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

requirement. Interventions on obesity are, therefore, not WDIs. Hernán and Taubman (2008, 513) conclude, "if the goal is to inform policy, it may be better to focus on modifiable lifestyle behaviors than on obesity itself."

Hernán and Taubman's preference for lifestyle variables is certainly well-motivated in certain respects. Lifestyle behaviors like physical exercise and diet control can be used as treatments in experiments, and interventions on treatments are designed to be non-ham-handed since there is no additional causal path going from a treatment assignment to the outcome of interest. Therefore, interventions on lifestyle behaviors are more likely to generate consistent results across different studies. In contrast, non-treatment variables like obesity are more likely to lead to "inconsistencies" between studies. In this respect, lifestyle behaviors are indeed better causal variables.

Nevertheless, their argument does not constitute a well-grounded objection to using obesity as a causal variable. Even if obesity is an inferior variable in certain respects, it has its own advantages, which should be taken into account in one's choice of causal variables, together with its disadvantages. First, obesity supports more general causal claims and causal estimations. This makes it possible to generalize results from various studies on the effects of diet control and exercise and predict the consequences of a broad range of specific behavioral causes. For example, if we want to predict the overall mortality trend in a country in the next few decades, it may be better to look for a few causes such as obesity, hypertension, and so on, instead of a larger number of behavioral causes.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Moreover, although it might be true that variables like diet and exercise are more useful for policymaking, these variables cannot fulfill our epistemic interests in the effects of biological characteristics. As a matter of fact, both the public and scientists are interested in the effects of obesity. In particular, when the goal is explanation, obesity or adiposity may be a better candidate for explaining health outcomes, since obesity is often a more proportional cause than lifestyle behaviors. This is also true for many other biological characteristics. Besides, it is worth mentioning that Pearl (2018) makes a similar argument as above by emphasizing the distinction between "policy-based" and "scientific" causation.

Even for purposes of policymaking, understanding the effects of obesity can still be important. If obesity is an important intermediate variable between lifestyle behaviors and deaths, specifying the causal pathways may help us achieve a better understanding and estimation of how lifestyle behaviors affect mortality. Finally, since the prevalence of obesity is easier to measure and monitor compared to the prevalence of unhealthy lifestyle behaviors, the former may turn out to be more useful for purposes of improving public health.

There is yet a more severe problem with Hernán and Taubman's reasoning. Granted that interventions on obesity may report "inconsistent" results between different studies, it is important to further identify the *source* of the inconsistency. If the source is in the inconsistency of the effect of obesity on mortality, then it probably means we should abandon obesity as a causal variable. But if the inconsistency can be entirely attributed to the ham-handedness of interventions, such an apparent "inconsistency" does not necessarily mean

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

obesity is a bad causal variable. In the latter case, what we should do is ask whether it is possible to identify CE(*Obesity*, *Mortality*) from these ham-handed interventions through some indirect methods.

Whether obesity has consistent effects on health outcomes under conceivable ideal interventions is a question Hernán and Taubman do not touch on. This question involves empirical issues about the physiology of human adipose tissue, especially how excessive white adipose tissue leads to diseases (this is an ongoing area of research; see Cypess 2022 for a review). If it turns out that causal pathways from white adipose tissue to health outcomes are significantly heterogeneous, this would suggest that CE(*Obesity*, *Mortality*) is ill-defined. I am not trying to do armchair physiology here. What I want to emphasize is that nothing in Hernán and Taubman's discussion supports that CE(*Obesity*, *Mortality*) is ill-defined.

Assuming that CE(*Obesity*, *Mortality*) is well-defined, there is still the question of whether we can reliably identify this effect from ham-handed interventions. Hernán and Taubman seem to assume that we cannot. But ideally speaking, there remains a possibility that we can identify CE(*Obesity*, *Mortality*) if we can eliminate or reduce the bias induced by the ham-handed path. For example, there may be available data on the direct effect of exercise (or diets) on mortality that does not go through obesity such that we can deduct the side effect. On the other hand, if there is a mediator Z on the ham-handed path from intervention I to outcome Y, as shown in figure 4(b), we can then adjust for Z to eliminate the bias.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Of course, these are merely theoretical possibilities; it may turn out that, in practice, there are no reliable ways to identify CE(*Obesity*, *Mortality*) from the data we have access to. In that case, Hernán and Taubman would still be justified in claiming that interventions on obesity are ill-defined. But this will not invalidate the key point I have made; namely, we should distinguish between the case in which CE(X, Y) is inconsistent and that in which different studies report apparently inconsistent results because interventions used in these studies are ham-handed (or nonideal). In the latter case, ham-handedness can simply be seen as a new source of bias in causal inference. Ham-handedness itself does not immediately suggest that the variable being manipulated is unsuitable for being a cause.

5. Conclusion

It is widely recognized that interventions invoked in a causal study must in some sense be "well-defined" and that using appropriate causal variables is essential for meeting this requirement. But there has been much debate on what sorts of interventions are well-defined and how considerations of this kind could provide guidance on variable choice. This paper contributes to the debate in the following respects:

1) This paper proposes a preliminary characterization of WDIs. Namely, for a causal study with CE(X, Y) as the effect of interest, an intervention on X with respect to Y is well-defined if and only if, 1) CE(X, Y) is well-defined under conceivable ideal interventions on X, and 2) the intervention is suitable for identifying CE(X, Y).

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

2) Interventions may be judged ambiguous for various reasons. An intervention is ambiguous when the effect of interest CE(*X*, *Y*) is ill-defined. CE(*X*, *Y*) is ill-defined when, for instance, *X* is a heterogeneous variable, or when *X* is ontologically controversial. In cases like these, we should be prepared to stop using *X* as a cause. On the other hand, an intervention may (appear to) be ambiguous when it is non-ideal (e.g., ham-handed). Here, we may continue to use *X* as a cause as long as we can reliably estimate CE(*X*, *Y*) from this non-ideal intervention (through some indirect methods).
3) To further illustrate the above points, a few typical examples of ambiguous interventions are studied. I want to stress again that the goal of these case studies is not to reach conclusive judgments on whether the variables discussed above are good causal variables. The goal is instead to reveal the complexities and subtleties surrounding the idea of WDIs and their roles in causal variable choice.

References

Chalupka, Krzysztof, Frederick Eberhardt, and Pietro Perona. 2016. "Multi-level Cause-effect Systems." *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR* 51:361-369.

————. 2017. "Causal Feature Learning: An Overview." *Behaviormetrika* 44(1):137-164.

Cypess, Aaron M. 2022. "Reassessing Human Adipose Tissue." *New England Journal of Medicine* 386(8):768-779.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Eberhardt, Frederick, and Richard Scheines. 2007. “Interventions and Causal Inference.” *Philosophy of Science* 74(5):981-995.

Glymour, Clark, and Madelyn R. Glymour. 2014. “Commentary: Race and Sex Are Causes.” *Epidemiology* 25(4):488-490.

Glymour, M. Maria, and Donna Spiegelman. 2017. “Evaluating Public Health Interventions: 5. Causal Inference in Public Health Research—Do Sex, Race, and Biological Factors Cause Health Outcomes?.” *American Journal of Public Health* 107(1):81-85.

Greene, Catherine. 2020. “Nomadic Concepts, Variable Choice, and the Social Sciences.” *Philosophy of the Social Sciences* 50(1):3-22.

Hamer, Dean, Brian Mustanski, Randall Sell, Stephanie A. Sanders, and Justin R. Garcia. 2021. “Comment on “Large-Scale GWAS Reveals Insights into the Genetic Architecture of Same-Sex Sexual Behavior.” *Science* 371(6536).

Hernán, Miguel A. 2016. “Does Water Kill? A Call for Less Casual Causal Inferences.” *Annals of Epidemiology* 26(10):674-680.

Hernán, Miguel A., and Sarah L. Taubman. 2008. “Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions.” *International Journal of Obesity* 32(3):S8-S14.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81(396):945-960.

————. 2008. “Causation and Race.” In *White Logic, White Methods: Racism and Methodology*, eds. Tukufu Zuberi, and Eduardo Bonilla-Silva, 93-109. New York: Rowman & Littlefield.

James, Michael, and Adam Burgos. 2020. “Race.” In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed. https://plato.stanford.edu/archives/sum2020/entries/race/.

Marcellesi, Alexandre. 2013. “Is Race a Cause?.” *Philosophy of Science* 80(5):650-659.

McCann, Kevin S. 2000. “The Diversity–Stability Debate.” *Nature* 405(6783):228-233.

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

————. 2018. “Does Obesity Shorten Life? Or Is It the Soda? On Non-manipulable Causes.” *Journal of Causal Inference* 6(2).

————. 2019. “On the Interpretation of *do(x)*.” *Journal of Causal Inference* 7(1).

Reiss, Julian. 2005. “Causal Instrumental Variables and Interventions.” *Philosophy of Science 72*(5):964-976.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66(5):688-701.

Schwartz, Sharon, Seth J. Prins, Ulka B. Campbell, and Nicolle M. Gatto. 2016. “Is the ‘Well-Defined Intervention Assumption’ Politically Conservative?.” *Social Science & Medicine* (1982)166:254-257.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press

Spirtes, Peter, Clark N. Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Spirtes, Peter, and Richard Scheines. 2004. "Causal Inference of Ambiguous Manipulations." *Philosophy of Science* 71(5):833–45.

VanderWeele, Tyler J. 2018. "On Well-defined Hypothetical Interventions in the Potential Outcomes Framework." *Epidemiology* 29(4):e24-e25.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

————. 2008. "Invariance, Modularity, and All That: Cartwright on Causation." In *Nancy Cartwright's philosophy of science*, eds. Stephan Hartmann, Carl Hoefer, and Luc Bovens, 198-237. New York: Routledge.

————. 2015. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91(2):303-347.

————. 2016. "The Problem of Variable Choice." *Synthese* 193(4):1047–72.

https://doi.org/10.1017/psa.2022.88 Published online by Cambridge University Press