Direct Causes and the Trouble with Soft Interventions

Frederick Eberhardt

Abstract

An interventionist account of causation characterizes causal relations in terms of changes resulting from particular interventions. I provide a new example of a causal relation for which there does not exist an intervention satisfying the common interventionist standard. I consider adaptations that would save this standard and describe their implications for an interventionist account of causation. No adaptation preserves all the aspects that make the interventionist account appealing. Part of the fallout is a clearer account of the difficulties in characterizing so-called "soft" interventions.

1 Introduction

James Woodward's (2003) account of causation characterizes causal relations in terms of *interventions*. On this account, speaking very roughly, for a variable x to be a cause of variable y, wiggling x must result in some change in (the probability of) y. Woodward's more circumspect description converts "wiggling" into a technical term. The result is an account that has a close connection to scientific practice and avoids dubious metaphysical baggage while still providing conceptual clarity of what it takes to stand in a causal relation.

In this article I present an example that challenges the specific commitments an interventionist endorses in characterizing a causal relation. The problem is neither entirely new, nor is it a problem the interventionist cannot avoid. It does, however, illustrate some of the details that may have gone unnoticed and forces any proponent of the interventionist account to be explicit about the assumptions she intends to make in order to avoid the uncomfortable implications described.

1.1 Interventionism

Woodward [2003, p. 55] provides the following definition of what it takes to be a *direct cause* on the interventionist account:

Definition 1 (Direct Cause (Woodward)) *A necessary and sufficient condition for x to be a direct cause of z with respect to some variable set V is that there be a possible intervention on x that will change z (or the probability distribution of z) when all other variables in V besides x and z are held fixed at some value by interventions.*[1]

Woodward provides other variations of this definition. I will return to these below, but for now it will suffice to note that I do not consider them to be substantially different for the present argument.

Definition 1 is most easily understood in terms of the representation of causal relations in so-called causal Bayes nets [Spirtes et al., 2000, Pearl, 2000]. In a causal Bayes net two variables are connected by an arrow whenever there is a direct causal relation between those variables. The resulting causal graph then gives rise to a probability distribution over the variables that satisfies the well-known Markov condition. The Markov condition states that each variable is probabilistically independent of its non-descendents given its parents in the graph. One way of understanding Definition 1 is that it provides a criterion consistent with the Markov condition for when a directed edge between two variables should be added to a causal graph (see Woodward [2003, p. 59]).

Several aspects of Woodward's definition are worth emphasizing:

1. The definition of a direct cause is relative to a set of variables V. While x may be a direct cause of z, i.e. $x \rightarrow z$, when we only consider the two variables $V = \{x, z\}$, it is possible that once we include the variable y in our considerations, the causal relation is in fact $x \rightarrow y \rightarrow z$, so x is no longer a direct cause relative to the set of variables $V = \{x, y, z\}$, only an indirect one. To avoid the requirement that all intermediary variables are included in V, the notion of a direct cause is relativized to V.

2. As the name of the account suggests, *interventions* play a special role. According to Woodward, one of the features of an intervention is that it is an exogenous influence that determines

the value of the intervened variable and makes the intervened variable independent of its normal causes (p. 96-98). This can be achieved by varying the variable as in a randomized experiment, or by locking the variable to a particular value. Although there are further details, for our purposes here it will suffice to note that the interventions Woodward proposes are of a particularly strong kind: they break the causal influences on the intervened variables and are therefore often referred to as "surgical" interventions. For the purpose of illustration, consider the effect of *drinking wine* on *heart disease*. One may worry that the correlation between *drinking wine* and *heart disease* is due to some confounder – a common cause of the two – such as *socio-economic status* (SES). But if one were to perform a controlled experiment in which participants were randomly assigned to a *wine drinking* or *no wine drinking* condition, then any influence of *SES* on *wine drinking* would be broken. The randomized controlled trial is a "surgical" intervention on *wine drinking*. The probability distribution arising from such an experiment is commonly referred to as a "manipulated distribution". One of the advantages of surgical interventions is that they can be performed without knowledge of the causal relations influencing the intervened variables. We need not know whether *SES* is in fact a cause of *wine-drinking* in order to perform the experiment. In Section 6 we will return to consider "softer" interventions that only nudge the intervened variable but may not break the influences of its other causes. While Definition 1 itself does not include an explicit restriction to surgical interventions (on x), Woodward's definition of an *intervention* in *Making Things Happen* only permits surgical interventions (IV.I2, p. 98).

3. Interventions in Definition 1 are existentially quantified, and modulated by the operator "possible". (In similar definitions Woodward has also used the term "hypothetical" or "plausible".) Woodward explains the motivation for the existential quantification as an explicitly weak requirement to permit interactive causes as direct causes. For example, a filled gas tank will only have an effect on the motor starting if the battery is also charged. If the battery is dead, then despite a full tank the motor will not start. So although the gas level has no effect on the motor starting when the battery is dead, it seems reasonable to consider the gas level to be a direct cause of the motor starting since it makes a difference for *some* setting of the

battery charge. The existential quantification captures such cases since it only requires that a change in x results in a change in z for *some* value assignment to the variables in $V \setminus \{x, z\}$.

A full justification and characterization of "possible" (interventions) is more difficult. Woodward discusses the issue in some detail in his Section 3.5. One motivation is to avoid the charge that the interventionist account of causation would otherwise appear inapplicable to causal relations in which an intervention does not seem feasible. For example, Woodward maintains that the gravitation of the Moon is a cause of the tides despite the fact that an intervention on the gravitation of the Moon does not appear feasible given our abilities (and is arguably physically impossible if everything else is supposed to be held fixed). For now we will rely on a suitably charitable reading and postpone the issue until Section 6.

Woodward may have intended Definition 1 to be couched in the context of additional background assumptions, although he is not explicit about them. We will consider such assumptions as we need them to handle the main example of this article.

2 Experimental Indistiguishability

Here, then, is the tricky case for the interventionist: Consider the two causal models T (triangle) and C (chain) over the binary variables $\{u, v, x, y, z\}$ in Figure 1. The variables x, y and z are observed, while u and v are unobserved, hence the dashed arrows. The models are identical except that in T the observed variable x is a direct cause of the observed variable z, i.e. $x \to z$, in addition to being an indirect cause of z via y. Table 2 specifies for each model all the parameters of the conditional probability distribution of each variable given its direct causes. Except for the (bold) parameters t_9 and t_{13} of the conditional probability of z given its causes, the parameterization of the two models are identical. Note that for model C the parameters

$$p(z|u, v, x = 1, y) = p(z|u, v, x = 0, y) = p(z|u, v, y) \quad \forall z, u, v, y,$$

so in model C the conditional distribution of z does not depend on x.

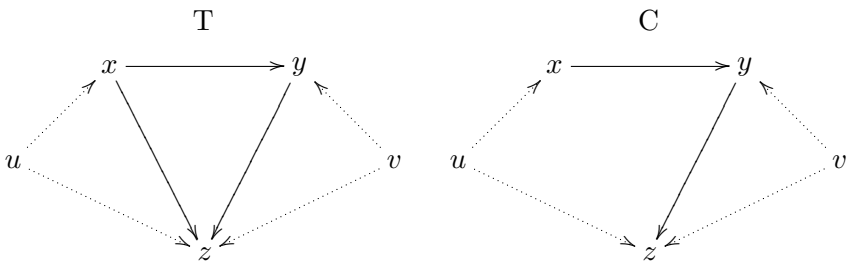


Figure 1: Model (T)riangle (left) and (C)hain (right). u and v are assumed to be unobserved variables, hence the dashed arrows.

What, according to a proponent of the interventionist account, justifies the direct cause $x \to z$ in model T?

The answer is not as straightforward as it may seem. Definition 1 depends on the set of variables we consider: If we take the perspective of a scientist who is only aware of the three observed variables x, y and z, then the set of variables under consideration is $V = \{x, y, z\}$. It follows from Definition 1 that x is a direct cause of z if and only if there is an intervention on x that results in a change in z while y is held fixed at 1 or at 0. It turns out that this is *not* the case for either model T or model C. In fact, if u and v are not observed then it can be verified that model T and C give rise to *exactly the same distribution* for

1. the passive observational distribution without interventions,
2. the manipulated distribution when only x is randomized,
3. the manipulated distribution when only y is randomized,
4. the manipulated distribution when only z is randomized,
5. the manipulated distribution when x and y are randomized simultaneously,
6. the manipulated distribution when x and z are randomized simultaneously, and
7. the manipulated distribution when y and z are randomized simultaneously.

Recall that identical joint distributions imply identical conditional and marginal distributions, so the fifth case includes as a conditional distribution the distribution when x is manipulated and y

parameter	conditional probability terms	T	C
t_1	$p(u = 1)$	0.3	0.3
t_2	$p(v = 1)$	0.4	0.4
t_3	$p(x = 1\|u = 1)$	0.8	0.8
t_4	$p(x = 1\|u = 0)$	0.2	0.2
t_5	$p(y = 1\|v = 1, x = 1)$	0.8	0.8
t_6	$p(y = 1\|v = 1, x = 0)$	0.8	0.8
t_7	$p(y = 1\|v = 0, x = 1)$	0.8	0.8
t_8	$p(y = 1\|v = 0, x = 0)$	0.2	0.2
$\mathbf{t_9}$	$\mathbf{p(z = 1\|u = 1, v = 1, x = 1, y = 1)}$	**0.65**	**0.8**
t_{10}	$p(z = 1\|u = 1, v = 1, x = 1, y = 0)$	0.8	0.8
t_{11}	$p(z = 1\|u = 1, v = 1, x = 0, y = 1)$	0.8	0.8
t_{12}	$p(z = 1\|u = 1, v = 1, x = 0, y = 0)$	0.8	0.8
$\mathbf{t_{13}}$	$\mathbf{p(z = 1\|u = 1, v = 0, x = 1, y = 1)}$	**0.9**	**0.8**
t_{14}	$p(z = 1\|u = 1, v = 0, x = 1, y = 0)$	0.8	0.8
t_{15}	$p(z = 1\|u = 1, v = 0, x = 0, y = 1)$	0.8	0.8
t_{16}	$p(z = 1\|u = 1, v = 0, x = 0, y = 0)$	0.8	0.8
t_{17}	$p(z = 1\|u = 0, v = 1, x = 1, y = 1)$	0.8	0.8
t_{18}	$p(z = 1\|u = 0, v = 1, x = 1, y = 0)$	0.8	0.8
t_{19}	$p(z = 1\|u = 0, v = 1, x = 0, y = 1)$	0.8	0.8
t_{20}	$p(z = 1\|u = 0, v = 1, x = 0, y = 0)$	0.8	0.8
t_{21}	$p(z = 1\|u = 0, v = 0, x = 1, y = 1)$	0.8	0.8
t_{22}	$p(z = 1\|u = 0, v = 0, x = 1, y = 0)$	0.2	0.2
t_{23}	$p(z = 1\|u = 0, v = 0, x = 0, y = 1)$	0.8	0.8
t_{24}	$p(z = 1\|u = 0, v = 0, x = 0, y = 0)$	0.2	0.2

Figure 2: Parameters of the two models in Figure 1. The differences between the models are shown in bold.

is held fixed at 0 (or 1) by an intervention. Given the graphical structures in Figure 1, the reader may note that for all these distributions the two models have exactly the same independence and dependence relations over $V = \{x, y, z\}$. The claim here, however, is stronger: The models have identical (manipulated) *distributions*.

The two models are thus *in principle indistinguishable* by passive observational data or by *any* (possibly simultaneous) surgical intervention on the observed variables. The models illustrate the most extreme form of experimental indistinguishability mentioned in Eberhardt [2012]. According to Definition 1, one must conclude that x is *not* a direct cause of z relative to $V = \{x, y, z\}$ in either T or (obviously) in C. Should, then, the arrow $x \rightarrow z$ in model T be omitted?

Here is a reason to think not: If instead of just the observed variables, we consider the enlarged set of variables $V^* = \{u, v, x, y, z\}$, then in an experiment that intervenes on x and holds the variables other than z fixed at $u = v = y = 1$, we have for model T

$$\begin{aligned}
& p_T(z = 1 | set(u = 1, v = 1, x = 1, y = 1)) \\
= \;& p_T(z = 1 | u = 1, v = 1, x = 1, y = 1) \\
= \;& t^T_9 \quad = \quad 0.65 \\
\neq \;& t^T_{11} \quad = \quad 0.8 \\
= \;& p_T(z = 1 | u = 1, v = 1, x = 0, y = 1) \\
= \;& p_T(z = 1 | set(u = 1, v = 1, x = 0, y = 1)),
\end{aligned}$$

where the $set(.)$-operator fixes the variables at particular values by intervention. But for model C

we have, as expected,

$$\begin{aligned}
& p_C(z = 1|set(u = 1, v = 1, x = 1, y = 1)) \\
= \;& p_C(z = 1|u = 1, v = 1, x = 1, y = 1) \\
= \;& t_9^C \\
= \;& 0.8 \\
= \;& t_{11}^C \\
= \;& p_C(z = 1|u = 1, v = 1, x = 0, y = 1) \\
= \;& p_C(z = 1|set(u = 1, v = 1, x = 0, y = 1))
\end{aligned}$$

We see that in model T the probability distribution of z changes depending on x, while all other variables are held fixed at some value. So by Definition 1, x is a direct cause of z relative to $V^* = \{u, v, x, y, z\}$ in T, but not in C.

So far there is nothing inconsistent with the interventionist account of causation: x is a direct cause of z relative to some V^*, but not relative to some other V. Nevertheless, it may be surprising that the direct causal effect of x on y in model T is not detectable by *any* surgical intervention on the observed variables $V = \{x, y, z\}$. The interventionist account is, among other things, supposed to support causal explanations and be closely related to how a scientist may go about establishing causal relations (see Woodward [2003, Sections 1.9 and 3.1.8]). How then should we react to this example?

On the one hand, it seems like the scientist is given all the tools she may desire – any randomized controlled trial on any set of the observed variables – but she is still in principle unable to detect the direct cause $x \to z$, *unless* she identifies the unobserved variables u and v first. On the other hand, for the set of variables V that she observes, it appears reasonable to claim that x is *not* a direct cause of z relative to $V = \{x, y, z\}$. After all, what would be the point of maintaining that x is a (direct) cause of z in model T? The above list of distributions that are identical for T and C shows that the direct causal effect only makes a difference to the (surgically) manipulated distributions once u and v are included in the set of variables under consideration. This was part of the reason in

the first place for relativizing the concept of direct cause to the set of variables under consideration. Note, however, the difference to the case mentioned in the first comment on Definition 1 above: There we had a shift from x as a direct cause of z, to x as an indirect cause of z. However, when we change the set of variables from V to V^* in the present case, model T exhibits a shift from x as an indirect cause of z, to x as an indirect and a direct cause of z, *of which neither causal path involves the variables u or v that were added into consideration.* The problem becomes most vivid in the manipulated distribution when x and y are subject to intervention (thereby breaking the pathway $x \to y \to z$): An entirely new causal connection between x and z appears – x becomes a direct cause of y – while apparently unrelated variables u and v are brought into view.

3 Causal Sufficiency

A natural first reaction to this example is to blame the unobserved variables. A similar example could not be constructed if *all* causal influences were observed.[2] But Woodward is careful not to endorse such a strong assumption. It would make the interventionist account of a direct cause inapplicable to most scientific contexts, since it is generally not the case that one observes *all* causal influences. Instead, Woodward explicitly endorses probabilistic causal connections with unobserved "error terms". These are common in the literature on structural equation models where a causal effect is determined by a function of its causes plus some unobserved disturbance. Often these disturbance terms are taken to be independent of one another, only influencing one variable each. In our models T and C, however, the unobserved variables u and v influence two variables each; they are so-called confounders or latent common causes. In the causal modeling literature the assumption of *causal sufficiency* is used to draw the line between independent disturbance terms and confounders: A set of variables is said to be causally sufficient if it contains all common causes of the set of variables, i.e. there are no latent confounders.

Glymour notes in his review of *Making Things Happen* that Woodward does not consider cases in which causal sufficiency is violated [Glymour, 2004]. Should Definition 1 consequently be read as implicitly referring to a causally sufficient set of variables?

I do not think so. Apart from the fact that it would undermine the relevance of the definition to

much of science where causal claims among causally insufficient sets of variables are investigated, there are reasons why the omission of causal sufficiency from Definition 1 may have been deliberate. The statistician Ronald Fisher is generally credited (or blamed?) for making randomized controlled trials the gold standard for causal discovery [Fisher, 1935]. One of the advantages that Fisher recognized was that the manipulation of the treatment variable according to a (causally) independent distribution made the treatment variable independent of its normal causes, including any unobserved confounders of the treatment and outcome. The same applies for Woodward's interventionist account: The surgical intervention on the potential cause breaks any confounding by unobserved variables (as we noted earlier in the case of *drinking wine*, *heart disease* and *SES*). The additional assumption of causal sufficiency therefore appears redundant.

More importantly, as can be seen when considering the graphical structures of model T and C in Figure 1, in the manipulated distribution when both x and y are subject to intervention, the causal influence of u on x and v on y are broken by the interventions. In this manipulated distribution u and v just function as independent "disturbance terms" on z. Thus, in the setting of Definition 1 that supposedly determines whether x is a direct cause of z, the set of variables $V = \{x, y, z\}$ is in fact causally sufficient. The bottom line is that there are not only good reasons why causal sufficiency would be a superfluous addition to Definition 1, but that even if it were added, it would not solve the problem exhibited by model T.

The unobserved variables are thus the wrong target for blame here. Instead, the problem arises due to an independence relation between x and z that is not implied by the causal structure. In the causal Bayes net literature such cases are known to occur when a particular assumption, known as *causal faithfulness*, is violated.

4 Faithfulness

Although Definition 1 does not include any explicit mention, causal faithfulness is a common assumption associated with causal discovery methods. Faithfulness states that all the independence relations in the probability distribution over the variables in V are a consequence of the Markov condition. One of the most common and well-understood violations of faithfulness occurs when

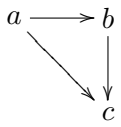


Figure 3: If the causal effects along the two paths from a to c cancel each other out exactly, then that is one way in which the model can exhibit a violation of the faithfulness condition.

there are two (or more) paths between variables that cancel each other out exactly (see e.g. Pearl [2000, p. 49]). Consider the three variables $W = \{a, b, c\}$ and suppose that they are causally related as shown in Figure 3. If each variable is determined by a linear function of its causal parents (plus some independent noise term) and the correlation between a and c due to the causal path $a \rightarrow b \rightarrow c$ cancels out exactly the correlation due to the direct effect of $a \rightarrow c$, then a and c will appear independent despite the fact that they are multiply causally related. Linearity plays no specific role other than that it is easy to understand. The following binary parameterization for the causal structure in Figure 3 results in a similar violation of faithfulness:[3]

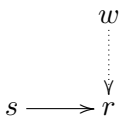
$p(a = 1)$	0.2
$p(b = 1 \mid a = 1)$	0.6875
$p(b = 1 \mid a = 0)$	0.2188
$p(c = 1 \mid a = 1, b = 1)$	0.6
$p(c = 1 \mid a = 1, b = 0)$	0.92
$p(c = 1 \mid a = 0, b = 1)$	0.2
$p(c = 1 \mid a = 0, b = 0)$	0.84

Variable a is (unconditionally) independent of c. Consequently, if b were not observed, then a and c would appear independent (violating faithfulness) in the passive observational distribution and in the manipulated distribution intervening on a, and would appear independent for an intervention on c (though not violating faithfulness in this case, obviously). *Unless* b is also observed, the causal paths from a to c are undetectable even with surgical interventions. Strevens [2008, p. 177] uses an example like this to argue that on Woodward's account causal relations can come into existence when the set of variables is expanded. Given that the model in Figure 3 exhibits a similar shift from no causal relation between a and c to a direct and indirect causal relation between the two variables depending on whether b is observed, should model T just be understood as a similar,

but more elaborate example of Figure 3 (or Strevens' example)?

No. While model T exhibits a violation of faithfulness[4], it is a violation of a different kind than the case of canceling pathways in Figure 3 or Strevens' example. First, unlike the model in Figure 3 and Strevens' example, model T does not violate faithfulness in the passive observational distribution. Second, when model T is subject to the surgical intervention on both x and y simultaneously (i.e. the case crucial to Definition 1), the causal effects $x \to y$, $v \to y$ and $u \to x$ are broken. Thus, in the true manipulated model of this intervention there are no two (or more) pathways between x and z whose causal effects could cancel each other out, there is only *one* path from x to z, namely the direct effect of $x \to z$ that remains. Still, this direct effect is not detectable by any surgical intervention or passive observation of $V = \{x, y, z\}$. The phenomenon of model T should also not be confused with the case of "single path unfaithfulness" that McDermott [1995, p. 531] describes. In that case an intermediary variable (between x and z) with at least three states would be required.

Model T results in particular discomfort for the interventionist account as all the desiderata of the account are fulfilled unambiguously: the intervention directly randomizes x independent of any other observed or unobserved variable (thus satisfying Woodward's response to Strevens' salty food example [Woodward, 2008]), the other observed variables are held fixed by an intervention (as required by Definition 1), there are no multiple pathways between x and z in the manipulated distribution relevant to the determination of the direct cause (in constrast to Figure 3 or Strevens' example of canceling pathways), and the variables that are brought into view – u and v – do not appear on pathways between the variables of interest (as in McDermott's case or the canceling pathways). This last point also makes any discussion of whether the situation can be saved by shifting the definition from *direct* to *contributing* causes redundant. One could just as well ask whether x is a *contributing cause* of z in the distribution in which x and y are manipulated – the problem remains.[5] Consequently, though also a violation of faithfulness, model T is not considered or addressed by Woodward's and Strevens' discussion of variable relativity [Strevens, 2007, Woodward, 2008, Strevens, 2008]. More specifically, Woodward could address Strevens' concern of canceling pathways without resolving the ambiguity arising from model T or could address the



$p(s=1)$	0.8
$p(w=1)$	0.5
$p(r=1\|s=1,w=1)$	0.2
$p(r=1\|s=1,w=0)$	0.8
$p(r=1\|s=0,w=1)$	0.8
$p(r=1\|s=0,w=0)$	0.2

Figure 4: A simple model with a noisy *xor*-parameterization. If w is not observed, then no surgical intervention on s or r can reveal the $s \to r$ edge.

problem of model T without handling canceling pathways. Or he could handle both by assuming faithfulness.

The violation of faithfulness that T exhibits is more similar to a model with a noisy *xor*-parameterization.[6] Consider the model in Figure 4 and its parameterization. If w is not observed, then s and r would appear independent in the passive observational distribution and in the manipulated distribution with an intervention on s, even though s is a direct cause of r. In Figure 4, however, the violation of faithfulness depends on the particular parameter of $p(w=1)=0.5$. For any other (non-extreme) value of that parameter, the direct cause of $s \to r$ is detectable. In contrast, due to its extra complexity, model T is not sensitive to a specific value of its parameters, even though it is sensitive to the relations among its parameters (see Appendix A). Nevertheless, model T and the model in Figure 4 share many similarities. In particular, in both cases the violation of faithfulness and the resulting undetectability of a direct causal effect is due to an averaging effect when summing over the unobserved variables. This is easily seen for the model in Figure 4:

$$\begin{aligned} p(r=1|s=1) &= \sum_w p(r=1|s=1,w)p(w) \\ &= 0.8 \times 0.5 + 0.2 \times 0.5 \\ &= \sum_w p(r=1|s=0,w)p(w) \quad = \quad p(r=1|s=0) \end{aligned}$$

Simiarly, obviously, for $r=0$, hence s is independent of r, i.e. $s \perp\!\!\!\perp r$.

The case is similar for model T when summing over the unobserved variables u and v.

Violations of faithfulness, at least those resulting from canceling paths, are widely discussed in the causal inference literature and familiar to proponents of the interventionist framework [Woodward, 2003, p. 49-50]. So it may seem surprising that none of the interventionist definitions of a causal relationship include the faithfulness condition. It would have provided a simple way to avoid all the problematic examples I have discussed so far. My guess is that the faithfulness assumption was deliberately omitted to avoid other undesirable consequences: First, in some cases violations of faithfulness are detectable once interventions are considered. For example, in the graph of Figure 3, if a, b and c are observed, then a and c would be unconditionally independent (due to the violation of faithfulness), but dependent given b. Given only passive observational data one may then incorrectly conclude that $a \to b \leftarrow c$ is the true model. But an intervention on b would easily resolve the confusion despite the unfaithfulness. Woodward uses an example like this to motivate his interventionist account [Woodward, 2003, p. 53].

Second, it is well known that deterministic causal relations trivially violate faithfulness. If Definition 1 depended on faithfulness, it would not apply to deterministic causal relations. Again, suitable surgical interventions can in many cases be used to identify even deterministic causal relations [Richardson et al., 2007, Glymour, 2007], so requiring faithfulness would seem like an unnecessarily strong restriction. For completeness I provide in Appendix B parameterizations of models T and C that are deterministic, but exhibit the same failure of detectability of the $x \to z$ edge for all surgical interventions on the observed variables. Obviously, these examples also show that the maximal change in the direct causal effect that is not detectable can be as high as 0 vs. 1. In general, the maximum causal effect that can be occluded depends on the other parameters of the model.

There are more reasons why the addition of faithfulness to Definition 1 would appear undesirable. For example, in causal relations with feedback, violations of faithfulness may be more plausible than the measure theoretic argument of Meek [1995] suggests. So, overall, assuming faithfulness does not seem like a promising route, though not an impossible one.

5 Population vs. Individual

In the previous section I pointed out that the undetectability of the $x \to z$ edge in model T given only $V = \{x, y, z\}$ is due to averaging that results from the marginalization of the unobserved variables u and v. Averaging, of course, requires more than one instance, and so it is tempting to hope that the problem is restricted to a population (or type) level account of causation and can be avoided at the token or unit level.

A little bit of background is necessary: In the philosophical literature on causation there is a common distinction that separates "actual causation" from "type level causation". The former is supposed to characterize the causal relations among events for one unit, while the latter considers causal relations at the population level. Details and criticism aside, very roughly the contrast is supposed to be something like "my smoking in the 80s in Europe caused my lung cancer in 2000 in the US" versus "smoking causes lung cancer". Among other things, the latter may still be true even if the former is not.

Woodward's account of causation discussed here follows Reichenbach and Suppes in considering causal relations at the population level. The suggestion for model T now goes as follows: When $V = \{x, y, z\}$ there may be an ambiguity about whether x is a direct cause of z at the population level. But for each individual unit in the population (think of a unit as a person in your sample population), is it not the case that either x directly caused z or not?

In order to answer this question, we need to establish what it takes to be a direct cause for a unit. In this context the natural way is to interpret Definition 1 at the unit level: *For unit* U, x is a direct cause of z with respect to some variable set V if and only if there is a possible intervention on *unit U's* variable x that will change *unit U's* variable z (or *unit U's* probability distribution of z), when all other *of unit U's* variables in V besides x and z are held fixed at some value by interventions. Various aspects would need to be spelled out in more detail (e.g. what are interventions on units? are the probabilities to be understood as propensities? etc.), but the gist is clear: Keep Definition 1, but restrict it to one unit.

This move will not resolve the problem: A single instance of manipulating a particular unit in the way suggested does not establish whether x is a unit-direct-cause of z (relative to V) because the

observed change in z (if any) of that unit may have resulted from an unobserved change of u and v in that unit. Consequently, to establish the direct cause within one unit, several manipulations *of the same unit* are required to establish the direct cause. But if one were to intervene repeatedly *on the same unit*, then each time the unobserved u and v can change for that unit as well, independently of the intervention. We are thus back where we started: We have reconstructed *for a single unit* the averaging effect that hides the $x \to z$ edge. The edge is undetectable in the multiple manipulations of the same unit given $V = \{x, y, z\}$. But the edge is easily detectable once $V^* = \{x, y, z, u, v\}$ is observed for *that unit*, just as at the population level.

One could insist, along the lines of more standard counterfactual analyses, that at the unit level, unobserved causes may not change when the unit is manipulated. Such a move weakens the importance of the intervention in the definition of direct cause and removes the close connection to causal discovery, as in practice there is no guarantee that unobserved causes remain fixed for a unit subject to intervention. Moreover, while addressing this violation of faithfulness, it fails to handle the case of canceling pathways (Figure 3). But the reader is free to insist on constraints that are not testable in principle.[7]

Of course, if the definition of a unit-direct-cause permits interventions that are dependent on the *un*observed variables u and v, then the problem of model T goes away.[8] But then one need not have resorted to the unit level, the problem would also disappear if one permitted such interventions in Definition 1 (at the population level). The main point is that a natural re-writing of Definition 1 for the unit level will not resolve the problem exhibited by model T. A more elaborate story would have to be told at the unit level and it is doubtful whether that story would then still be in the spirit of Woodward's original definition. Definition 1 includes an epistemological component of how we come to learn the causal relations, and without a move to some form of population level, there is currently no epistemology of actual (unit) causal relations that considers unobserved variables. Some philosophers may want to resort to an exclusively metaphysical truth about the underlying causal relations at the unit level, but the importance of the connection to epistemology in Woodward's project cannot be denied.

6 Interventions

I have considered three general approaches of how an interventionist may respond to the ambiguity in Definition 1 illustrated by model T. All of these either do not work or look unappealing. There are other routes one could take, some of which are mentioned in Section 7 below, but they have a much more ad hoc nature. Instead, the interventionist could just bite the bullet and maintain that x is not a direct cause of z in model T relative to $V = \{x, y, z\}$, but is a direct cause relative to $V^* = \{x, y, z, u, v\}$. Leaving aside the (at least metaphysical) concern resulting from direct causes popping in and out of existence, it seems as though for each specific set of observed variables the practical consequences are well-defined. Such a pragmatic argument hinges on an assumption that when only $V = \{x, y, z\}$ is observed, the $x \to z$ edge is in fact completely undetectable, not just undetectable by surgical interventions.

This assumption is false. There exist *soft*, rather than surgical, interventions such that models C and T are distinguishable even if only the variables in $V = \{x, y, z\}$ are observed. For example, a soft intervention on y that changes the parameter $t_5 = p(y = 1|v = 1, x = 1)$ from 0.8 to 0.85 would make models T and C distinguishable (see Appendix C). No other variable would need to be intervened on. In fact, this generalizes: For all models that violate faithfulness (including Strevens' case of canceling pathways), there exist soft interventions that make the causal relations detectable without having to enlarge the variable set (see Appendix C).

What are soft interventions? In contrast to surgical interventions they are a weaker form of intervention. For example, a surgical intervention on the variable "income" would set the income of the participants in the experiment independently of its normal causes. But alternatively, one could also consider the effect of an intervention that only adds, say, \$5,000 to each participant's income. In that case, the variable "income" is still influenced by its normal causes (education, etc.), but the intervention adds an additional "nudge". Such an intervention is often referred to as a "soft" intervention. To give a maximally general account of an intervention, one could only require that an intervention change the conditional probability distribution of the intervened variable given its normal causes. A surgical intervention makes the intervened variable independent of its normal causes, while a soft intervention is an intervention that is not surgical.

In model T a soft intervention on y that only changes the parameter $t_5 = p(y = 1|v = 1, x = 1)$ from 0.8 to 0.85 leaves the dependencies of y on its causes (x and v) intact in the manipulated distribution. A surgical intervention would break the causal influences and make y independent of its causes, replacing all the parameters of y's conditional probability distribution with a single parameter for $p^*(y = 1)$ of the manipulation.

Woodward does not consider soft interventions. He permits a wide scope of "possible" interventions, including, as I noted in the introduction, interventions that are arguably physically impossible, but all his interventions are supposed to be surgical (see Woodward [2003, IV.I2, p. 98]). Nevertheless, one could extend Definition 1 to include soft interventions, though one would have to decide whether *any* (combination of) soft interventions on any variable is permitted. Only a soft intervention on y or z, and only particular soft interventions at that, would distinguish model T and model C when $V = \{x, y, z\}$. A soft intervention on x or a soft intervention on t_6 would not be sufficient (see Appendix A).

This points to the crux with soft interventions: While it is easy to define a soft intervention on a particular model mathematically, the success of a correct implementation of a soft intervention is not always testable given the observed data. Although it is not unique in its ability to distinguish between model T and model C when $V = \{x, y, z\}$, the soft intervention on y that only changes the parameter $t_5 = p(y = 1|v = 1, x = 1)$ from 0.8 to 0.85 is sufficient to distinguish the two models, and illustrates the difficulty of implementing soft interventions: Suppose as a scientist we only observe $V = \{x, y, z\}$, but are aware that there may be unmeasured latent variables. How would we implement the soft intervention that changes t_5 in the way described? We can see when x equals 1, but we do not know when $v = 1$. Nevertheless, the intervention is not supposed to affect $t_7 = p(y = 1|v = 0, x = 1)$, at least not in exactly the same way as it affects t_5. Without observing v then, the success of the soft intervention is not directly verifiable. But there is no logical reason that prevents a justification of the assumption that such a soft intervention is successful in specific cases. For example, we may have domain knowledge about how particular chemicals work and that they can manipulate specific interactions (such as that between x and v on y) even if we do not know whether v is present or what value it takes. In such cases we may have reason to

believe that a soft intervention of the kind described can be implemented successfully. One can imagine other circumstances where suitable domain knowledge and appropriate tools are available. The implementation of soft interventions cannot be specified in terms of the intervened variables, because the target of a soft intervention is a particular parameters of a variable rather than the whole variable. The parameters may not be known, not even how many there are.

This limitation constitutes good reason to be cautious about basing an entire account of causation on soft interventions. But the difficulty of implementation in general does not undermine the point that it is sometimes possible to use soft interventions to discover causal relations when surgical interventions are insufficient. Moreover, the difficulty found with soft interventions suggests a closer look at the conditions for a surgical intervention. Woodward reminds the reader explicitly that an intervention makes the intervened variable independent not only of its observed causes, but also of its unobserved causes [Woodward, 2008]. Above I loosely spoke of a surgical intervention as manipulating a variable, while a soft intervention manipulates a parameter of the variable. But a surgical intervention as Woodward describes it, amounts to manipulating *all* the parameters of that variable. Also, just like the case for soft interventions, without observing v there is no way of directly verifying whether a *surgical* intervention on y successfully made y independent of v. The success of such an intervention rests on an assumption about the randomizing device and a sufficiently large sample. While not the same, the difficulties in the implementation of soft interventions are not entirely foreign to the implementation of surgical interventions. At the very least, a more explicit account is required of what aspects of an intervention should be verifiable, before clean lines can be drawn between the two types. Woodward says nothing about the testability of his criteria for interventions (see the parts of Definition IV on p. 98 in Woodward [2003]) and this testability is, as has been noted in Baumgartner [2012], not obvious.

I see the situation then as follows: Due to the difficulties in providing a general implementation technique, it does not seem promising to include soft interventions in the definition of direct cause. But I do think it is very plausible that there are specific circumstances where soft interventions of the form described are well justified and successfully implementable. This implies that model T and model C are sometimes distinguishable on the basis of soft interventions even when just

$V = \{x, y, z\}$ are observed. Thus, a direct cause that is not covered by Definition 1 is sometimes detectable, and so one *cannot* simply retreat to the pragmatic position that x is a direct cause of z relative to $V^* = \{x, y, z, u, v\}$ in model T, but is not a direct cause of z relative to $V = \{x, y, z\}$ in model T. The ardent interventionist can now disallow soft interventions[9] (citing, for example, some of the reasons mentioned above) or can hope that there is no nifty scientist who provides a concrete example of a discovery of a direct cause as I describe, thus leaving Definition 1 practically unchallenged. Neither option seems desirable.

7 Implications for Causal Discovery

In light of the discussion in this paper it is evident that Definition 1 is not sufficiently precise to form part of a basis for a causal discovery algorithm. Unless cases such as model T are excluded by additional assumptions, there will be cases where the detection of direct causal effects will not depend only on the surgical interventions that are possible.

Apart from faithfulness, the problem of model T could be resolved by supplementing Definition 1 with a requirement that causal relations have a particular parametric form. For example, one cannot parameterize model T with linear causal relations without making the $x \to z$ edge detectable for some surgical intervention on a subset of the variables in V. This holds quite generally: For linear relations Hyttinen et al. [2012] show that even when faithfulness is not assumed, all causal relations among a set of observed but causally insufficient variables can be detected by a set of experiments each intervening surgically on a (different) single variable. However, linearity constitutes, like faithfulness, a strong assumption about the causal relations among the set of variables. It is known to be violated in many real causal relations, and there are causal discovery procedures, especially ones involving interventions, that do not depend on it.

Alternatively, one could modify the assumption of causal sufficiency such that it must hold not only for the crucial distribution of Definition 1, but also for the passive observational distribution (which is clearly not the case for model T). But one should then consider just how much of science the resulting definition would not apply to: I have so far described the unobserved variables u and v in model T as if they were well-defined causal variables that – if they were observed –

could be subject to intervention. However, often latent variables are used as a catch-all for various background effects that result in confounding. For many inference procedures that permit latent confounding, there may not be a commitment that the latent variable is a particularly nicely defined variable that can be subject to intervention. In econometrics, latent confounding often just represents any type of correlation in the error variables. Consequently, the skepticism that has been raised concerning the possibility of performing all the interventions required in the standard interventionist account, applies in much stronger form to potential interventions on variables we currently do not observe. Consequently, a demand that in principle one can, so to speak, always zoom out far enough to capture all unobserved variables, is too strong.

I expect cases like model T to raise interesting issues for causal discovery from data sets that do not share the same set of variables, so-called overlapping data sets. For example, suppose that the true underlying causal structure has the form of model T, but one research group collects data (possibly using surgical interventions) over the variables $V = \{x, y, z\}$, and another research group collects data (also using surgical interventions) over the variables $W = \{u, x, z\}$. The first research group will not detect the $x \to z$ edge, while the second will. How should they now combine their findings? It seems that their findings conflict, but in fact we know that each group found exactly what they should, given the underlying true model. The appropriate inference principles to combine the results must still be worked out.

A further aspect that is neglected by Definition 1 is that causal relations may involve feedback: What should count as a direct effect of x on z if z also has an effect on itself? Ordinarily, such feedback relations are represented in terms of time series or differential equations. The detection of feedback from data, especially if it involves "self-loops", is known to be difficult. Depending on how exactly one characterizes the feedback, the notion of direct cause may change. Hyttinen et al. [2012] discuss this issue in some detail for the linear case in their Section 2.3, and proceed to use a standardized notion of direct cause that includes the self-loops on the non-intervened variable, but no feedback via other observed variables.

8 Conclusion

I have argued that Woodward's interventionist account of a direct cause runs into difficulties with particular cases of violations of faithfulness that I believe have not been discussed and analyzed in this way before. Although I have focused on Woodward's definition, as stated here in Definition 1, I believe that if anything, it is among the least problematic among (interventionist) definitions of 'cause'. The argument presented here can be adapted easily to apply to other purely interventionist definitions, as well as to anthropocentric definitions of cause that in addition to interventions, build on the presence of an agent [Menzies and Price, 1993]. Regularity accounts of cause are known to have problems with violations of faithfulness and do not consider causally insufficient sets of variables, while mechanistic accounts have no systematic commitment in the first place about how the causal relations manifest themselves in measured data. So all these other definitions are in my view substantially vaguer with regard to their commitments and subject to additional criticism. Definition 1 is in that sense pleasantly clear and widely applicable.

I repeat that the argument I have given does not imply an inconsistency in Woodward's definition. My challenge on the basis of model T and model C can be avoided by a requirement that the causal models be faithful, or by any of the other modifications I have pointed to. For any responses of this type, it only behooves a proponent of the interventionist account to be more explicit, and complete the commitments they subscribe to in defining a direct cause (or a cause or a contributing cause). I have suggested that none of these additional commitments are particularly desirable because they come at the expense of the virtues that make the interventionist account so appealing. They require stronger assumptions that are known to be violated or they require knowledge about unobserved variables that undermines the relevance of the account to scientific practice. But if one leaves the details of the connection to causal discovery aside, one may just accept that along the edges most concepts have counterexamples.

A Constraints for Models T and C

I consider the general constraints on the parameterization that two models of the structure of T and C must satisfy in order to be indistinguishable for a passive observation and all surgical interventions on the observed variables. In the following T and C refer to models with the respective structures in Figure 1, rather than the specific parameterizations listed in Table 2, and I use the notation $p(\mathbf{A}|\mathbf{B}||\mathbf{C})$ to refer to the probability of the variables in $\mathbf{A}$ conditional on the variables in $\mathbf{B}$ in the distribution in which the variables in $\mathbf{C}$ have been subject to a surgical intervention.

To remain indistinguishable when $V = \{x, y, z\}$, model T and model C must be identical for the following seven distributions over the observed variables:

1. the passive observational distribution:

$$P(X,Y,Z) = \sum_{uv} P(U)P(V)P(X|U)P(Y|V,X)P(Z|U,V,X,Y)$$

2. the manipulated distribution[10] with an intervention on X

$$\begin{aligned} P(Y,Z|X||X) &= \sum_{uv} P(U)P(V)P(Y|V,X||X)P(Z|U,V,X,Y||X) \\ &= \sum_{uv} P(U)P(V)P(Y|V,X)P(Z|U,V,X,Y) \end{aligned}$$

To illustrate the implied constraints, I substitute the distribution parameters from Table 2

for this particular case. It can be done analogously for all seven distributions.

$$
\begin{aligned}
P(y=1,z=1|x=1||x=1) &= t_1t_2t_5t_9 + (1-t_1)t_2t_5t_{17} + t_1(1-t_2)t_7t_{13} + (1-t_1)(1-t_2)t_7t_{21} \\
P(y=1,z=0|x=1||x=1) &= t_1t_2t_5(1-t_9) + (1-t_1)t_2t_5(1-t_{17}) \\
&\quad + t_1(1-t_2)t_7(1-t_{13}) + (1-t_1)(1-t_2)t_7(1-t_{21}) \\
P(y=0,z=1|x=1||x=1) &= t_1t_2(1-t_5)t_{10} + (1-t_1)t_2(1-t_5)t_{18} \\
&\quad + t_1(1-t_2)(1-t_7)t_{14} + (1-t_1)(1-t_2)(1-t_7)t_{22} \\
P(y=0,z=0|x=1||x=1) &= t_1t_2(1-t_5)(1-t_{10}) + (1-t_1)t_2(1-t_5)(1-t_{18}) \\
&\quad + t_1(1-t_2)(1-t_7)(1-t_{14}) + (1-t_1)(1-t_2)(1-t_7)(1-t_{22}) \\
P(y=1,z=1|x=0||x=0) &= t_1t_2t_6t_{11} + (1-t_1)t_2t_6t_{19} \\
&\quad + t_1(1-t_2)t_8t_{15} + (1-t_1)(1-t_2)t_8t_{23} \\
P(y=1,z=0|x=0||x=0) &= t_1t_2t_6(1-t_{11}) + (1-t_1)t_2t_6(1-t_{19}) \\
&\quad + t_1(1-t_2)t_8(1-t_{15}) + (1-t_1)(1-t_2)t_8(1-t_{23}) \\
P(y=0,z=1|x=0||x=0) &= t_1t_2(1-t_6)t_{12} + (1-t_1)t_2(1-t_6)t_{20} \\
&\quad + t_1(1-t_2)(1-t_8)t_{16} + (1-t_1)(1-t_2)(1-t_8)t_{24} \\
P(y=0,z=0|x=0||x=0) &= t_1t_2(1-t_6)(1-t_{12}) + (1-t_1)t_2(1-t_6)(1-t_{20}) \\
&\quad + t_1(1-t_2)(1-t_8)(1-t_{16}) + (1-t_1)(1-t_2)(1-t_8)(1-t_{24})
\end{aligned}
$$

3. the manipulated distribution with an intervention on Y

$$
\begin{aligned}
P(X,Z|Y||Y) &= \sum_{uv} P(U)P(V)P(X|U,Y||Y)P(Z|U,V,X,Y||Y) \\
&= \sum_{uv} P(U)P(V)P(X|U)P(Z|U,V,X,Y)
\end{aligned}
$$

4. the manipulated distribution with an intervention on Z (since this distribution does not involve the parameters specifying $p(z|u,v,x,y)$ that distinguish the models, these equations are trivially satisfied by T and C)

$$
P(X,Y|Z||Z) = \sum_{uv} P(U)P(V)P(X|U)P(Y|V,X)
$$

5. the manipulated distribution with an intervention on X and Y simultaneously

$$\begin{aligned} P(Z|X,Y||X,Y) &= \sum_{uv} P(U)P(V)P(Z|U,V,X,Y||X,Y) \\ &= \sum_{uv} P(U)P(V)P(Z|U,V,X,Y) \end{aligned}$$

6. the manipulated distribution with an intervention on X and Z simultaneously (since this distribution does not involve the parameters specifying $p(z|u,v,x,y)$ that distinguish the models, these equations are trivially satisfied by T and C)

$$\begin{aligned} P(Y|X,Z||X,Z) &= \sum_{uv} P(U)P(V)P(Y|U,V,X,Z||X,Z) \\ &= \sum_{v} P(V)P(Y|V,X) \end{aligned}$$

7. the manipulated distribution with an intervention on Y and Z simultaneously (since this distribution does not involve the parameters specifying $p(z|u,v,x,y)$ that distinguish the models, these equations are trivially satisfied by T and C)

$$\begin{aligned} P(X|Y,Z||Y,Z) &= \sum_{uv} P(U)P(V)P(X|U,V,Y,Z||Y,Z) \\ &= \sum_{u} P(U)P(X|U) \end{aligned}$$

In addition, in order to establish the relevant causal effects, both models must satisfy the following inequalities. The bold font indicates (at least one way) in which the parameterizations of T and C in Table 2 satisfy the inequalities.

1. to make u a cause of x:

$$\mathbf{t_3} \neq \mathbf{t_4}$$

2. to make x and v causes of y:

$$((t_5 \neq t_7) \vee (\mathbf{t_6} \neq \mathbf{t_8})) \quad \wedge \quad ((t_5 \neq t_6) \vee (\mathbf{t_7} \neq \mathbf{t_8}))$$

3. to make u, v and y a cause of z:

$$\begin{aligned}
&((\mathbf{t_9} \neq \mathbf{t_{17}}) \vee (t_{10} \neq t_{18}) \vee (t_{11} \neq t_{19}) \vee (t_{12} \neq t_{20}) \vee (t_{13} \neq t_{21}) \vee (t_{14} \neq t_{22}) \vee (t_{15} \neq t_{23}) \vee (t_{16} \neq t_{24})) \\
\wedge&((\mathbf{t_9} \neq \mathbf{t_{13}}) \vee (t_{10} \neq t_{14}) \vee (t_{11} \neq t_{15}) \vee (t_{12} \neq t_{16}) \vee (t_{17} \neq t_{21}) \vee (t_{18} \neq t_{22}) \vee (t_{19} \neq t_{23}) \vee (t_{20} \neq t_{24})) \\
\wedge&((\mathbf{t_9} \neq \mathbf{t_{10}}) \vee (t_{11} \neq t_{12}) \vee (t_{13} \neq t_{14}) \vee (t_{15} \neq t_{16}) \vee (t_{17} \neq t_{18}) \vee (t_{19} \neq t_{20}) \vee (t_{21} \neq t_{22}) \vee (t_{23} \neq t_{24}))
\end{aligned}$$

Model T must in addition make x a cause of z by satisfying the following inequality:

$$\begin{aligned}
&(\mathbf{t_9} \neq \mathbf{t_{11}}) \vee (\mathbf{t_{13}} \neq \mathbf{t_{15}}) \vee (t_{17} \neq t_{19}) \vee (t_{21} \neq t_{23}) \\
\vee \quad &(t_{10} \neq t_{12}) \vee (t_{14} \neq t_{16}) \vee (t_{18} \neq t_{20}) \vee (t_{22} \neq t_{24})
\end{aligned} \tag{1}$$

while model C must satisfy its negations, i.e. all the parameter pairs must be equal.

Since model T must satisfy at least one disjunct of Constraint (1), while C must satisfy its negation, one can easily detect the distributional constraints from the list 1-7 above that will not be trivially satisfied. All such quantities contain either only parameters from the first line, or only parameters from the second line of Constraint (1). I will focus only on the satisfaction of disjuncts from the first line, the case for the second line is exactly analogous.

In the most general case model T differs from model C by satisfying every disjunct in Constraint (1), and we can write the parameters as $t_9 = t_{11} + d_1$, $t_{17} = t_{19} + d_2$, $t_{13} = t_{15} + d_3$, and $t_{21} = t_{23} + d_4$ for non-zero $d_1, \ldots, d_4$. There are seven distributional quantities containing the parameters t_9, t_{13}, t_{17} and t_{21}, giving rise to the following four independent constraints if models T and C are to be indistinguishable for a passive observation and all surgical interventions on the

observed variables:

$$\begin{aligned}
t_1t_2t_3t_5d_1 + (1-t_1)t_2t_4t_5d_2 + t_1(1-t_2)t_3t_7d_3 + (1-t_1)(1-t_2)t_4t_7d_4 &= 0 \\
t_1t_2t_5d_1 + (1-t_1)t_2t_5d_2 + t_1(1-t_2)t_7d_3 + (1-t_1)(1-t_2)t_7d_4 &= 0 \\
t_1t_2t_3d_1 + (1-t_1)t_2t_4d_2 + t_1(1-t_2)t_3d_3 + (1-t_1)(1-t_2)t_4d_4 &= 0 \\
t_1t_2d_1 + (1-t_1)t_2d_2 + t_1(1-t_2)d_3 + (1-t_1)(1-t_2)d_4 &= 0
\end{aligned}$$

Solving these constraints implies that a model T must satisfy the following constraints on its parameters

$$\begin{aligned}
t_5 &= t_7 \\
t_{11} &= t_9 - (d_3(-1+t_2)/t_2) \\
t_{15} &= t_{13} - d_3 \\
t_{19} &= t_{17} - (d_4(-1+t_2)/t_2) \\
t_{23} &= t_{21} - d_4
\end{aligned} \tag{2}$$

where d_3 and d_4 can be chosen freely as long as at least one of them is non-zero and the resulting quantities remain probabilities. An analogous set of constraints results when the difference between models T and C results from disjuncts in the second line of Constraint (1). These are non-trivial algebraic constraints on the parameter space, which, following Meek [1995], implies that their solution space has measure zero compared to arbitrary parameterizations of a model with a structure like T.

Similarly, these constraints can be used to construct a parameterization for a model T that is indistinguishable from a parameterized model C, as long as the parameterization of C also respects the $t_5 = t_7$ constraint (or $t_6 = t_8$). In particular, the parameterization of T in Table 2 is constructed from the parameterization of C in that table using $d_3 = 0.1$ and $d_4 = 0$.

B Deterministic Parameterizations of T and C

Deterministic parameterizations of the two models in Figure 1 that are indistinguishable for a passive observation and any surgical intervention on the observed variables.

parameter	conditional probability terms	T	C
t_1	$p(u=1)$	0.5	0.5
t_2	$p(v=1)$	0.5	0.5
t_3	$p(x=1\|u=1)$	0	0
t_4	$p(x=1\|u=0)$	1	1
t_5	$p(y=1\|v=1,x=1)$	1	1
t_6	$p(y=1\|v=1,x=0)$	1	1
t_7	$p(y=1\|v=0,x=1)$	1	1
t_8	$p(y=1\|v=0,x=0)$	0	0
t_9	$p(z=1\|u=1,v=1,x=1,y=1)$	1	0
t_{10}	$p(z=1\|u=1,v=1,x=1,y=0)$	1	1
t_{11}	$p(z=1\|u=1,v=1,x=0,y=1)$	0	0
t_{12}	$p(z=1\|u=1,v=1,x=0,y=0)$	1	1
t_{13}	$p(z=1\|u=1,v=0,x=1,y=1)$	0	1
t_{14}	$p(z=1\|u=1,v=0,x=1,y=0)$	1	1
t_{15}	$p(z=1\|u=1,v=0,x=0,y=1)$	1	1
t_{16}	$p(z=1\|u=1,v=0,x=0,y=0)$	1	1
t_{17}	$p(z=1\|u=0,v=1,x=1,y=1)$	1	1
t_{18}	$p(z=1\|u=0,v=1,x=1,y=0)$	1	1
t_{19}	$p(z=1\|u=0,v=1,x=0,y=1)$	1	1
t_{20}	$p(z=1\|u=0,v=1,x=0,y=0)$	1	1
t_{21}	$p(z=1\|u=0,v=0,x=1,y=1)$	1	1
t_{22}	$p(z=1\|u=0,v=0,x=1,y=0)$	1	1
t_{23}	$p(z=1\|u=0,v=0,x=0,y=1)$	1	1
t_{24}	$p(z=1\|u=0,v=0,x=0,y=0)$	1	1

Note that if the latent variables u and v are supposed to be non-extreme, then only $u = v = 0.5$ are possible values.

I do not find the deterministic case particularly enlightening. Moreover, it is well known that deterministic causal relations are often more difficult to discover than probabilistic ones. In that sense I think that the examples of parameterizations for model T and C in Table 2 with purely positive distributions provide a much stronger case.

C Soft Interventions

Note that the constraints in (2) do not contain the parameters t_3 or t_4 which could be influenced by a soft intervention on x, hence a soft intervention on x is not going to distinguish between models T and C.

A soft intervention on y that changes t_5 will break the first equality in (2), thus the models become distinguishable. In particular, if t_5 is changed from 0.8 to 0.85 by a soft intervention on y in both models T and C, then in the resulting manipulated distribution, we will have

$$\begin{aligned} p^*_T(x = y = z = 1) &= 0.24856 \\ \text{vs.} \quad p^*_C(x = y = z = 1) &= 0.24928 \end{aligned}$$

which is not a rounding error.

Note that t_6 does not feature in the constraints in (2), so a soft intervention that changes it in both T and C, will not distinguish between the two models.

Lastly, as Meek [1995] showed, violations of faithfulness occur for particular constellations of the parameters that make up the distribution. A violation of faithfulness always constitutes a non-trivial algebraic constraint on these parameters. Since soft interventions influence individual parameters, they can be (in principle, leaving the mentioned concerns of implementation aside) used to break these algebraic constraints. Thus, soft interventions are in general sufficient to make unfaithful models faithful. Once faithfulness is achieved, the causal relations can be detected as usual. In the case of canceling pathways with unobserved intermediary variables (Figure 3), the soft intervention must occur on the final variable.

Notes

[1]Note that I have exchanged y for z from the original formulation to reduce confusion in the application of the definition in the subsequent discussion.

[2]For the close reader, I literally mean "all" here, i.e. even noise terms. As will be seen in Figure 4 below, cases similar to model T are possible when particular unobserved noise terms are permitted.

[3]This parameterization was constructed by marginalizing over a noisy-or model where b is a negative (inhibiting) cause of c.

[4]In its standard form, faithfulness only refers to the passive observational distribution, and since the models in Figure 1 do not exhibit any (conditional) independencies in the passive observational distribution, they do not violate the standard formulation of faithfulness. However, it is natural to extend faithfulness to apply to all manipulated graphs and their interventional distributions as well. Model T clearly violates this stronger version of faithfulness, since it leaves x and z independent in the distribution where x and y are simultaneously subject to an intervention, even though x is a direct cause of z (as determined by the intervention on the full causal graph including u and v). As with violations of standard faithfulness, model T exhibits a particular constellation of parameters that can be characterized by an algebraic constraint on the parameters. Meek's measure theoretic argument that such constellations only occur with measure zero (see Theorem 7 in Meek [1995]) can be similarly applied here (see Appendix A).

[5]I am grateful to an anonymous reviewer for making this proposal. I hope this clarifies why such a move will not work, at least not with the definition of 'contributing cause' that Woodward gives in Woodward [2008, p. 209].

[6]I am grateful to Dominik Janzing for pointing this out. The example is similar to the "matching pennies" game, only with one coin flip unobserved.

[7]I am grateful to a reviewer for alerting me to this route and at the same time supplying the reasons why it does not sound promising.

[8]Standard discussions of actual causation do not consider unobserved variables, so the problem exhibited by model T does not arise.

[9]The interventionist may want to be cautious not to throw out the baby with the bathwater, since instrumental variables have formally the same structure as soft interventions and are widely used as a causal discovery tool.

[10]I condition on the intervened variable(s) in order to avoid having to specify a particular intervention distribution.

Acknowledgements

Many people, including many who are not proponents of the interventionist account of causation, have reacted with discomfort or at least some surprise to my presentation of model T and C. I have benefitted enormously from their reactions and discussions with them. In particular, I would like to thank (in alphabetical order) Clark Glymour, Dominik Janzing, Conor Mayo-Wilson, Richard Scheines, Peter Spirtes, Jim Woodward and Jiji Zhang. The models are a development of ones that are indistinguishable by single interventions only, which I worked on in a different context with Antti Hyttinen and Patrik Hoyer. I would also like to thank five anonymous reviewers who pressed me to distinguish more explicitly the case presented here from traditional violations of faithfulness

discussed in the literature. This research was supported by a grant from the James S. McDonnell Foundation.

References

M. Baumgartner. The logical form of interventionism. *Philosophia*, 40(4):751–761, 2012.

F. Eberhardt. Experimental indistinguishability of causal structures, 2012. URL `http://philsci-archive.pitt.edu/9511/`. Forthcoming in Proceedings of PSA 2012.

R. A. Fisher. *The design of experiments*. Hafner, 1935.

C. Glymour. Review of James Woodward, *Making Things Happen: A Theory of Causal Explanation*. *British Journal for Philosophy of Science*, 55:779–790, 2004.

C. Glymour. Learning the structure of deterministic systems. In A. Gopnik and L. Schulz, editors, *Causal learning: Psychology, philosophy, computation*. Oxford University Press, 2007.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439, 2012.

M. McDermott. Redundant causation. *British Journal for Philosophy of Science*, 46:523–544, 1995.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *UAI 1995*, pages 411–418, 1995.

P. Menzies and H. Price. Causation as a secondary quality. *British Journal for Philosophy of Science*, 44:187–203, 1993.

J. Pearl. *Causality*. Oxford University Press, 2000.

T. Richardson, L. Schulz, and A. Gopnik. Data-mining probabilists or experimental determinists? A dialogue on the principles underlying causal learning in children. In A. Gopnik and L. Schulz, editors, *Causal Learning: Psychology, Philosophy, Computation*, pages 208–230. Oxford University Press, 2007.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2nd edition, 2000.

M. Strevens. Review of Woodward, *Making Things Happen*. *Philosophy and Phenomenological Research*, 74(1):233–249, 2007.

M. Strevens. Comments on Woodward, *Making Things Happen*. *Philosophy and Phenomenological Research*, 77(1):171–192, 2008.

J. Woodward. *Making Things Happen*. Oxford University Press, 2003.

J. Woodward. Response to Strevens. *Philosophy and Phenomenological Research*, 77(1):193–212, 2008.