# Direct Cause

**Frederick Eberhardt** fde@cmu.edu
*Department of Philosophy*
*Carnegie Mellon University*
*Pittsburgh, Pennsylvania, USA*

**Editor:**

## Abstract

An interventionist account of causation characterizes causal relations in terms of changes resulting from particular interventions. We provide an example of a causal relation for which there does not exist an intervention satisfying the common interventionist standard. We 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.

## 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 $y$ (possibly only in probability). Woodard's more circumspect description converts "wiggling" into a technical term. The interventionist account has provided a pragmatic middle road, avoiding dubious metaphysical baggage while still providing a conceptual clarity of what it takes to stand in a causal relation. By building on the representation of causal relations in terms of graphical models, the interventionist account also connects to causal inference procedures used in the sciences.

In this article we 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 we describe.

## 2. 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. We briefly return to these at the end of the article. For now it will suffice to note that we 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: First, 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, that 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 $V$.

Second, 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 fixing 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.* We refer to the probability distribution arising from such an experiment 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* or not 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 does not include an explicit restriction to surgical interventions, Woodward

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

does not discuss soft interventions and his definition of an intervention in “Making Things Happen” only permits surgical interventions (IV.I2, p. 98).

Lastly, 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.

## 3. 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. $x \rightarrow z$, 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 $x$.

*What, according to a proponent of the interventionist account, justifies the direct cause $x \rightarrow z$ 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

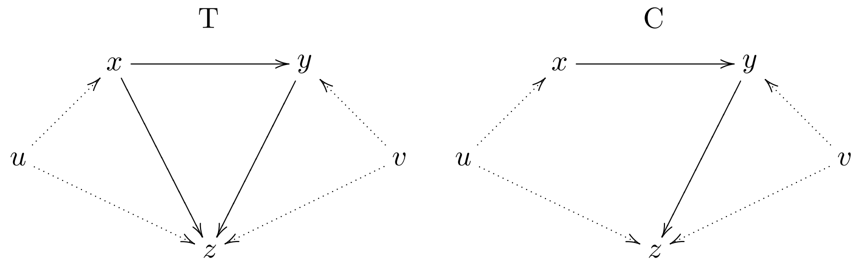


Figure 1: Model ($T$)riangle (left) and ($C$)hain (right). $u$ and $v$ are assumed to be unobserved variables, hence the dashed arrows.

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

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

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 and independently,
6. the manipulated distribution when $x$ and $z$ are randomized simultaneously and independently, and
7. the manipulated distribution when $y$ and $z$ are randomized simultaneously and independently.

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 $y$ 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. 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 $x \to z$ in model $T$ be omitted?

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 $T$

$$\begin{aligned}
& p_T(z=1|set(u=1, v=1, x=1, y=1)) \\
&= \; p_T(z=1|u=1, v=1, x=1, y=1) \\
&= \; t_9^T \quad = \quad 0.65 \\
&\neq \; t_{11}^T \quad = \quad 0.8 \\
&= \; p_T(z=1|u=1, v=1, x=0, y=1) \\
&= \; p_T(z=1|set(u=1, v=1, x=0, y=1)),
\end{aligned}$$

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

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

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 be a pragmatic account, supporting causal explanations and 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, 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 $x \rightarrow z$, *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 from a pragmatic perspective 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, as we suggested before, 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, that unlike the shift from direct to indirect cause that we discussed in the context of the relativization in Definition 1, when we change the set of variables from $V$ to $V^*$, 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.

## 4. 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

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

of a direct cause imapplicable 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 an effect is 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* draws 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.

As Glymour notes in his review of "Making things happen", Woodward does not consider cases in which causal sufficiency is violated (Glymour, 2004), as is the case when only $V = \{x, y, z\}$ are observed in models $T$ and $C$. Should Definition 1 consequently be read as implicitly referring to a causally sufficient set of variables?

We do not think so. Apart from the fact that much of science, which the interventionist after all wants to relate to, investigates causal claims among causally insufficient sets of variables, 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. Moreover, 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. Thus, in the setting of Definition 1 that supposedly determines whether $x$ is a cause of $z$, the set of variables $V = \{x, y, z\}$ is in fact causally sufficient. In this manipulated distribution $u$ and $v$ just function as independent "disturbance terms" on $z$. 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. The problem has more to do with an independence relation between $x$ and $z$ that is not implied by the causal structure. In the causal Bayes nets literature such cases are known to occur when a particular assumption, known as *causal faithfulness*, is violated.

## 5. 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

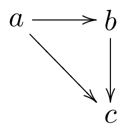


Figure 3: If the causal effect along the two paths from $a$ to $c$ cancel each other out exactly, then the model violates the faithfulness condition.

of the Markov condition. Violations of faithfulness are well known to result in situations where particular causal relations cannot be detected. One of the most common and well-understood violations of faithfulness occurs when there are two paths between variables that cancel each other out exactly. 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 |
| $p(b = 1\|a = 1)$ | 0.6875 |
| $p(b = 1\|a = 0)$ | 0.2188 |
| $p(c = 1\|a = 1, b = 1)$ | 0.6 |
| $p(c = 1\|a = 1, b = 0)$ | 0.92 |
| $p(c = 1\|a = 0, b = 1)$ | 0.2 |
| $p(c = 1\|a = 0, b = 0)$ | 0.84 |

Variable $a$ will be (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.

In the causal discovery literature it is standard practice, and often quite reasonable, to assume that the faithfulness assumption is not violated: As is intuitive from the linear example involving canceling paths, a violation of faithfulness depends on a very specific constellation of parameters to render two causally connected variables independent. The situation is similar in the binary case. This intuition is supported by a measure-theoretic result that shows that with respect to Lebesgue measure over the set of linearly independent parameters of a multinomial distribution that is Markov to the graph, a violation of faithfulness has measure zero (see Theorem 7 in Meek (1995)).

[3] This parameterization was constructed by marginalizing over a noisy-or model where $b$ is a negative (inhibiting) cause of $c$ (Hyttinen et al., 2010).

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 $T$ also just be understood as a similar, but more elaborate example of a violation of faithfulness?

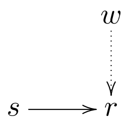
On the one hand, the cases are similar: Model $T$ constitutes a relatively straightforward violation of faithfulness. In its standard form, faithfulness only refers to the passive observational distribution, and since the models in Figure 1 do not exhibit any (conditional) independencies in the passive 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, and Meek's measure theoretic argument can be similarly applied to show that such a violation has measure zero (see Appendix). From a measure theoretic point of view then, faithfulness can be strengthened so that cases such as model $T$ count as a violation, while violations of faithfulness remain measure-zero events. Definition 1 could thus be restricted to such "strongly" faithful causal relations, thereby avoiding $T$ by excluding a space of parameters that has measure zero.

On the other hand, despite its measure theoretic rarity, there are a few additional points worth noting about $T$. First, for any parameterization of the "chain"-model $C$ that satisfies the constraint $(t_5 = t_7) \vee (t_6 = t_8)$, it is easy to construct a parameterization of model $T$ that will result in identical passive observational and manipulated distributions (as listed in Section 3) over the observed variables $V = \{x, y, z\}$. Moreover, in practice such a constraint need only hold approximately since the test to distinguish the two is only based on a finite sample size. If one considered the possibility of an additional third unobserved variable between $x$ and $y$, then there are even more ways in which the parameters can satisfy constraints to make the models indistinguishable from one another. So while formally an event of measure zero, there are many situations where one may think one has discovered a model with the structure among the observable variables resembling that of model $C$, but where in fact there is another model that additionally has an $x \rightarrow z$ edge which is in principle undetectable given any passive observation or surgical intervention on the observed variables.[4]

Second, note that the violation of faithfulness exhibited by $T$ is not a case of canceling paths. In particular, note that when model $T$ is subject to a surgical intervention on both $x$ and $y$ simultaneously, then the $x \rightarrow y$, the $v \rightarrow y$ and the $u \rightarrow x$ edge are broken. Thus, in this manipulated model there is only one path from $x$ to $z$, namely the direct effect of $x \rightarrow z$ that remains. Still, this direct effect is not detectable by any surgical intervention or passive observation.[5]

[4] We currently do not know the general conditions of when this occurs in arbitrary structures, nor do we have a more precise measure theoretic account.

[5] Also, $T$ is not a case of "single path unfaithfulness", as described by McDermott (1995, p. 531). In that case an intermediary variable with at least three states is needed.



| $p(s=1)$ | 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 $s \rightarrow r$ edge.

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 $s \rightarrow r$ is detectable. In constrast, 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). 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 $s \perp\!\!\!\perp r$.

The case is similar for model $T$ when summing over the unobserved variables $u$ and $v$.[7]

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 we have discussed so

[6] I am grateful to Dominik Janzing for pointing this out.

[7] Despite the fact that the source of the phenomenon lies in this averaging effect, one should not misunderstand the problem that model $T$ illustrates as one pertaining exclusively to population level causal claims. One could always interpret the probabilities as individual propensities. But, of course, the discovery of individual causal propensities is unclear if one does not have some assumption about how propensities relate to some population of instances.

far. Our guess is that the faithfulness assumption was deliberately omitted to avoid other problems. First, in many 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 $a \rightarrow b \leftarrow c$ is the true model. But an intervention on $b$ would easily resolve the confusion despite the unfaithfulness. Woodward uses an almost identical example to motivate his interventionist account (Woodward, 2003, p. 53). Thus, one could maintain the view that sometimes interventions can be used to overcome violations of faithfulness, and for those cases where they do not – such as model $T$, or when $b$ is not observed in the model in Figure 3, or for the model in Figure 4 – there is a pragmatic reason *not* to infer a causal relation, since the causal relation makes no difference under any surgical intervention or passive observation of the observed variables.

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 we provide in the appendix parameterizations of models $T$ and $C$ that are deterministic, but exhibit the same failure of detectability of the $x \rightarrow z$ 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 the Definition 1 would appear undesirable. For example, in causal relations with feedback, violations of faithfulness may be more plausible than the measure theoretic argument suggests. So, overall, this does not seem like a promising route, though not an impossible one.

Alternatively, one can, of course, insist that the interventionist should handle the examples presented here analogously to how she handles the violations of faithfulness due to canceling pathways. Namely, pragmatism dictates that $x$ is not a direct cause of $z$ in model $T$ relative to $V = \{x, y, z\}$, because the direct influence makes no difference under any surgical intervention on the observed variables. In Figure 3, $a$ is also not judged a direct cause of $c$ when $b$ is unobserved. Once the unobserved variables are included, then the direct causal relations change, since now they do make a detectable difference. Such a pragmatic argument hinges on an assumption of undetectability of the direct cause by interventions. It is worth analyzing the details of such an assumption, especially for models, like $T$, that exhibit a violation of faithfulness.

## 6. Interventions

In the discussion of Definition 1 in Section 2 we noted Woodward's insistence on a wide scope of "possible" interventions to characterize causal relations. Woodward is quite explicit in accepting that many of the interventions required to detect causal relations may not be practically feasible (the effect of the gravitational field of the Moon on the tides is a case in point). To require *some* intervention to determine a direct causal relation thus does not constitute a requirement of being actually able to perform the intervention, but is something akin to a *conceptual* criterion of what it takes to be a direct cause. It is not entirely clear how wide the scope of interventions considered in Definition 1 is supposed to be, but it appears uncontroversial that it should include all physically possible interventions Reutlinger (2012).

Such a reading of the scope of interventions may also explain why Woodward only considers "surgical" interventions. In contrast to surgical interventions there are weaker forms 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.

Woodward does not discuss soft interventions, so it is not clear whether soft interventions are to be excluded from Definition 1. While there are many cases where a surgical intervention is not *practically feasible*, but a soft intervention is, it may still be within the scope of *physical possibility* that for any physically possible soft intervention, there is also a physically possible surgical intervention on the same variables. Would that make the omission of soft interventions in Definition 1 innocuous?

Again the answer is not straightforward and depends on the exact reading of Definition 1: On the one hand, if only $V = \{x, y, z\}$ of model $T$ are observed, then there is no soft intervention that changes $x$ that will change $z$ (or the probability distribution of $z$) when all other variables in $V$ besides $x$ and $z$ (i.e. $y$) are held fixed at some value by a (surgical) intervention. So on this reading, the omission of soft interventions is innocuous if the physical possibility of surgical interventions is taken to be suitably wide. A proponent of the interventionist account could thus maintain that Definition 1 need not encompass soft interventions, since the same scope can be achieved with surgical ones, and the fact that the $x \rightarrow z$ edge in $T$ is not detected, is not a bug, but a pragmatic virtue of the account, since for all surgical interventions on the observed variables, the $x \rightarrow z$ edge cannot be detected anyway.

On the other hand, however, *there exist soft interventions, for example, on $y$, such that models $C$ and $T$ are distinguishable even if only the variables $V = \{x, y, z\}$ are observed.* For example, a soft intervention 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). No other variable would need to be intervened on.

Definition 1 does not consider soft interventions on $y$ to detect whether $x$ is a direct cause of $z$. So it is not clear what an interventionist who subscribes to Definition 1 would say to this case. The dilemma is that using a soft intervention on $y$, a scientist is in principle able to determine the presence of the $x \rightarrow z$ edge in model $T$ (while variables $u$ and $v$ remain unobserved), but cannot do the same using just surgical interventions.

Unfortunately, broadening Definition 1 to encompass soft interventions on *any* observable variable entails its own problems. Soft interventions have extremely weak requirements that can make their implementation difficult. Changing the parameter $t_5 = p(y = 1|v = 1, x = 1)$ from 0.8 to 0.85 constitutes a soft intervention, but it is easier described than implemented. We noted that one of the advantages of a surgical intervention is that it can be implemented without knowing the causal context of the intervened variable. The same does not apply to soft interventions in general. Moreover, it is *not* the case that *all* soft interventions on $y$ are sufficient to detect the $x \to z$ edge. In particular, a soft intervention that changes only $t_6$ does not distinguish models $T$ and $C$.

The issue becomes more poignant when one considers what it means to perform the *same* intervention on two different causal models. Models $T$ and $C$ have the virtue (vice?) that in terms of the conditional probability distribution of $y$ given its parents, they are identical. But consider an intervention on $z$. In the case of a *surgical* intervention on $z$, it is trivial to say when such an intervention in model $T$ would be the same as in model $C$: The intervention would have to set the variable $z$ to the same fixed value or according to the same randomizing distribution, so that the marginal manipulated distribution of $z$ is the same no matter whether the underlying causal model is $T$ or $C$. For a soft intervention there need not be such demand, since a soft intervention is not supposed to make the intervened variable independent of its normal causes. By construction, models $T$ and $C$ are such that a soft intervention on $y$ that changes only $t_5$ does result in the same marginal manipulated distribution over $y$ in both $T$ and $C$. But for a soft intervention on $z$, the same need not hold: While it is possible to perform a soft intervention that changes $t_9$ but not $t_{11}$ in model $T$, the same is impossible in model $C$, since $t_9$ and $t_{11}$ are not only equal, but *identical*; they must change together. In general then, the demand for identical marginal distributions over the intervened variables may be considered an unnecessarily strong standard for performing the "same" soft intervention on different underlying causal models. Unlike surgical interventions that manipulate parameters in bulk, soft interventions can manipulate parameters individually. But parameters, even how many there are, may well be unknown.

One could retreat and claim that soft interventions are ill-defined, at least for the purpose of establishing causal relations. For the general case there seems to be some truth to that. But the earlier example of a soft intervention on the variable "income" shows that there certainly appear to exist soft interventions that can be implemented and are easily understood even though the full causal structure was not described.[8] Similarly, unless one wants to deny the possibility of soft interventions for causal discovery in general[9], the case here of the specific soft intervention on $y$, changing parameter $t_5$, is one of the least controversial soft interventions, since the parameterization of the target of the intervention, i.e. $y$, is the same in the two models $T$ and $C$ that are to be distinguished.[10]

[8] Of course, it is possible that other tacit assumptions were at play.

[9] Recall that instrumental variables have formally the same structure as soft interventions.

[10] We note that there also exist soft interventions that can be used to detect causal relations in other cases of violations of faithfulness, such as canceling pathways (Figure 3) or deterministic causal relations. In some of those cases, however, there may not be soft interventions as simple as the one on $t_5$ here. We will not pursue the issue here.

We have referred to the requirement in Definition 1, that there exist *some* suitable intervention, as a *conceptual* requirement of a direct cause, because it does not require that a scientist should actually be able to perform the intervention that would allow her to detect the direct cause. However, we have shown that there exist cases, such as model $T$ versus model $C$, where surgical interventions do not identify a direct causal relation, which can be detected by a soft intervention. According to a literal reading of Definition 1 such a soft intervention should not be considered relevant, because it intervenes on the wrong variable (and is soft). So we are left with a causal relation detected by a soft intervention that is not permitted by Definition 1. Its conceptual appeal looks tarnished.

One could re-write Definition 1 to include soft interventions on other variables, thereby including the causal relations only detectable by soft interventions. But this imports the difficulties of how to make sense of the implementation of many soft interventions in general when the causal structure is not already known. Thus broadening Definition 1 in order to preserve its *conceptual* appeal of tracking the detectability of causes comes at the cost of weakening its *pragmatic* appeal.

The upshot is that soft interventions break apart the virtues of Definition 1: One can save the *conceptual* criterion of being a direct cause, but one loses the *pragmatic* appeal of the definition that connects to discovery procedures in science. Or one can maintain the pragmatically cleaner stance that denies the direct cause from $x$ to $z$ (relative to $V = \{x, y, z\}$) in model $T$ and one then may have to acknowledge that a nifty scientist can show the presence of a direct causal effect that supposedly could not exist.

Another way of looking at it is to acknowledge that Woodward's existential quantification over "possible" interventions is doing an enormous amount of work skirting a line between a mathematically well defined search space for causal relations, and characterizing conceptually satisfying interventions.

## 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 only depend on the surgical interventions that are possible.

Apart from faithfulness, model $T$ could be excluded 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 $x \rightarrow z$ edge detectable for some surgical intervention on a subset of the variables in $V$ (Eberhardt et al., 2010). 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 actual cases of 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: We 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. A demand that in principle one can, so to speak, always zoom out far enough to capture all unobserved variables, may be too strong, and a suitable soft intervention on $z$ may be much easier to perform, or at least more plausible.

We 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 $x \rightarrow z$ 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

We have argued that Woodward's interventionist account of a direct cause runs into difficulties with particular cases of violations of faithfulness that we believe have not been analyzed in this way before. Although we have focused on Woodward's definition, as stated here in Definition 1, we 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 seem to presuppose that one knows everything already anyway. So all these other definitions are in our view substantially vaguer with regard to their commitments and subject to additional criticism. Definition 1 is in that sense pleasantly clear and widely applicable.

We repeat that the argument we have stated does not imply an inconsistency in Woodward's definition. Our 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 we 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). We 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. 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.

## Acknowledgements

Many people, including many who are not proponents of the interventist 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. This research was supported by a grant from the James S. McDonnell Foundation.

## Appendix A. Constraints for Models $T$ and $C$

We 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. We thus now use $T$ and $C$ to refer to models with the respective structures in Figure 1, rather than the specific parameterizations listed in Table 2, and we 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.

Model $T$ and model $C$ must be identical for the following distributions over the observed variables.

1. the passive observational distribution:

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

2. the manipulated distribution[11] 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, we substitute the parameters for this particular case. It should be done analogously for all the other seven distributions.

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

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 not relevant)

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

5. the manipulated distribution with an intervention on $X,Y$

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

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

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

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

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

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

In addition, both models must satisfy the following inequalities. The bold font indicates (at least one way) how 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. We 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 $d_4 = 0$.

## Appendix B. Soft Interventions

Note that the constraints in (2) do not contain the parameters $t_3$ or $t_4$ which would 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.

Lastly, 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.

## Appendix C. 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.

We 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 we think that the examples of parameterizations for model $T$ and $C$ in Table 2 with purely positive distributions provide a much stronger case.

## References

F. Eberhardt, P. O. Hoyer, and R. Scheines. Combining experiments to discover linear cyclic models with latent variables. In *AISTATS '10*, 2010.

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. Causal discovery for linear cyclic models with latent variables. In *PGM '10*, 2010.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. Submitted, available at
`http://www.cs.helsinki.fi/u/ajhyttin/hyttinen2011sub.pdf`, 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.

A. Reutlinger. Getting rid of interventions. *Studies in the History and Philosophy of Science. Part C*, 2012.

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, 2 edition, 2000.

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