The Transitivity and Asymmetry of Actual Causation*

Sander Beckers[1] and Joost Vennekens[2]

[1] Cornell University
[2] KU Leuven, Dept. Computer Science @ Campus De Nayer

The problem of actual causation has received a lot of attention in both the philosophy and the AI literature. In a nutshell, the problem is to define when event X is deemed to have caused event Y in the context of a particular story. Typically, it is assumed that this story unfolds according to a given set of causal laws. Coming up with a suitable definition for this concept of actual causation has proven to be quite difficult.

In this paper, we attempt to both explain why this is so difficult, and to offer a method for potentially solving the problem. We do so by focusing on two natural, yet mutually incompatible intuitions, namely that actual causation should be both transitive and asymmetric.

Following the approach of the seminal work by Halpern and Pearl [2], we make use of structural models as our formal tool.

Example 1. An assassin-in-training is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim.

We can represent this example by means of a structural model that consists of five boolean random variables and two equations that relate these variables:

$$VictimDies := TraineeHits \vee SupervisorHits$$
$$SupervisorHits := SupervisorShoots$$
$$TraineeHits := TraineeShoots$$
$$SupervisorShoots := \neg TraineeShoots$$

The story told in the example corresponds to the following assignment of values to these variables:

$$TraineeShoots = TraineeHits = VictimDies = true;$$
$$SupervisorShoots = SupervisorHits = false.$$

In his seminal work, Lewis [3] proposed to define actual causation as the transitive closure of counterfactual dependency. In the case of the above example,

* The full version of this paper was published as [1].

$VictimDies$ is counterfactually dependent on $TraineeHits$ (in the context of the story, in which $SupervisorHits$ is false, it is the case that if trainee's bullet hadn't hit the victim, the victim would not have died) and $TraineeHits$ is counterfactually dependent on $TraineeShoots$ (if he hadn't shot, then his bullet wouldn't have hit the victim). Therefore, by transitivity, victim's death was caused by trainee's shooting, even though it does not counterfactually depend on it (if trainee would not have shot, then victim would still have died).

The transitivity of actual causation invoked by Lewis seems intuitively plausible. Indeed, if X caused Y and Y in turn caused Z, then surely it is fair to call also X a cause of Z. However, several convincing counterexamples have emerged, such as the following example by [4].

Example 2. Terrorist, who is right-handed, must push a detonator button at noon to set off a bomb. Shortly before noon, he is bitten by a dog on his right hand. Unable to use his right hand, he pushes the detonator with his left hand at noon. The bomb duly explodes.

$$Bomb := LeftHand \vee RightHand$$
$$LeftHand := DogBite$$
$$RightHand := \neg DogBite$$

Clearly, the dog bite caused the terrorist to use his left hand, and the terrorist's use of his left hand caused the explosion. Nevertheless, it seems too farfetched to call the dog bite a cause for the explosion, as transitivity would mandate. Therefore, this example demonstrates that the intuition of transitivity has its limitations.

In this paper, we claim that the limiting factor is the intuition of asymmetry, i.e., that we should never consider X to have caused Y if, at the same time, we would also have considered $\neg X$ (if $\neg X$ had been the case instead of X) as a cause of the same Y. In the case of the above example, we see that, in the context where the dog would not have bitten the terrorist, there would be counterfactual dependence of $Bomb$ on $RightHand$ and of $RightHand$ on $\neg DogBite$. Therefore, Lewis' account would count both $DogBite$ and $\neg DogBite$ as a cause for the same event $Bomb$, which would violate asymmetry.

In the full version of this paper, we argue for the hypothesis that asymmetry is the limiting factor on the transitivity of actual causation, i.e., that actual causation is precisely as transitive as it can be without violating asymmetry.

References

1. Sander Beckers and Joost Vennekens. The transitivity and asymmetry of actual causation. *Ergo, an Open Access Journal of Philosophy*, 2017.
2. J. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British Journal for the Philosophy of Science*, 56:843–87, 2005.
3. D. Lewis. Causation. *Journal of Philosophy*, 70:113–126, 1973.
4. M. McDermott. Redundant causation. *The British Journal for the Philosophy of Science*, 46(4):523–544, 1995.

The Transitivity and Asymmetry of Actual Causation

Abstract

The counterfactual tradition to defining actual causation has come a long way since Lewis started it off. However there are still important open problems that need to be solved. One of them is the (in)transitivity of causation. Endorsing transitivity was a major source of trouble for the approach taken by Lewis, which is why currently most approaches reject it. But transitivity has never lost its appeal, and there is a large literature devoted to understanding why this is so. Starting from a survey of this work, we will develop a formal analysis of transitivity and the problems it poses for causation. This analysis provides us with a sufficient condition for causation to be transitive, a sufficient condition for dependence to be necessary for causation, and several characterisations of the transitivity of dependence. Finally, we show how this analysis leads naturally to several conditions a definition of causation should satisfy, and use those to suggest a new definition of causation.

1 Introduction

Causal modelling has become ubiquitous in Artificial Intelligence circles, and is gaining popularity in other fields as well. An unsolved problem in this context is how to define actual causation, i.e., when should we say that one event caused another? Ever since Lewis (1973) first analyzed this problem in terms of counterfactual dependence forty years ago, philosophers and researchers from the Artifical Intelligence community alike have been trying to improve on his attempt at cracking this causal nut. The seminal work of Halpern and Pearl (2005) has led to the structural equations framework becoming the most important language to deal with this problem.

The currently most prominent approaches to defining actual causation are those within the counterfactual dependence tradition, which started with Lewis (1973). All of these approaches take as their starting point the assumption that counterfactual dependence is sufficient for causation, but not necessary (Hitchcock, 2001; Woodward, 2003; Hall, 2004, 2007; Halpern and Pearl, 2005; Halpern, 2016; Weslake, 2015). That dependence is sufficient is usually accepted simply as a fundamental principle underlying causation. That it is not necessary, on the other hand, is usually defended by pointing to intuitively strong counterexamples. Lewis (1973) forms an important exception to this rule, as he defends the lack of necessity by invoking a principle as well, namely that causation is transitive: causation is transitive, dependence is not, therefore there can be causation without dependence.

The first strategy, that of offering counterexamples, has proven most successful. There are two reasons for this. First, almost everyone besides Lewis rejects the transitivity of causation. Second, there are counterexamples to the necessity of dependence that have nothing to do with transitivity. Despite its success, this strategy has to date not offered a general insight into precisely when or why the transitivity of causation breaks down. Although a substantial number of authors have addressed the problem of transitivity, none of them offers a generally sufficient condition for causation to be transitive (McDermott, 1995; Hall, 2000, 2004; Hitchcock, 2001; Sartorio, 2005; Halpern and Pearl, 2005; Halpern, 2015; Paul and Hall, 2013). A recent discussion by Halpern (2015) does formulate several sufficient and necessary conditions for transitivity, however those apply only to cases where there is dependence.

The main contribution of this paper is to offer a principled explanation of why transitivity should be rejected as a general condition, while also offering conditions under which it should be satisfied. Specifically, we will explain both why the transitivity of causation has a strong appeal, and why there are nevertheless convincing counterexamples to accepting it. We do so by an appeal to the principle that causation is *asymmetrical*: an event is a cause only if its absence would not have been a cause.[1] Accepting this prin-

[1]Note that usually the asymmetry of a relation $R(x, y)$ is in-

ciple leads the way to an analysis of causation as a transitive relation compromised by asymmetry. This analysis provides us with a sufficient condition for causation to be transitive, a sufficient condition for dependence to be necessary, and several sufficient and necessary conditions for dependence to be transitive. Finally, we use this analysis to suggest a new definition of causation. The starting point for our analysis consists of a detailed overview of the literature on this topic.

Our story will be incomplete: we ignore one important type of example, Late Preemption, that highlights the temporal aspects of causation. This caveat will not undermine the current discussion, because it stands orthogonal to the issue of transitivity. In fact one can integrate this temporal aspect into our analysis to give a more refined definition of causation, which is what we aim to do in future work.

The next section introduces the structural equations framework. Sections 3 offers the relevant background, and presents a survey of the literature on transitivity. This leads us to suggest a first condition any definition of causation should satisfy, in Section 4. Section 5 introduces the concept of *contributing*, which leads the way to a sufficient condition for the transitivity of causation. We discuss the asymmetry of causation in Section 6 and show how it can be combined with transitivity to form an elegant explanation of all aspects here discussed.

2 Structural Equations Modelling

We briefly introduce a simple version of structural equations modelling, which is the most popular formal language used to represent causal models. In general, structural equations allow functional dependencies between continuous variables, or discrete variables with possibly an infinite domain. However, the actual causation literature typically considers only examples made up of discrete variables with a finite domain, and propositional formulas. Further, in the majority of cases the variables are Boolean. This is why we restrict attention to those kinds of models. For a detailed introduction, see (Pearl, 2000).

A structural model consists of a set of *endogenous* variables $\vec{V}$, a set of *exogenous* variables $\vec{U}$, and a causal model M. Although we only consider models with Boolean variables, we should point out that the results we will present can easily be generalized to allow for multi-valued variables as well. We explain this below.

A model M is a set of *structural equations* so that there is exactly one equation for each variable $V_i \in \vec{V}$. An equation takes the form $V_i := \phi$, where ϕ is a propositional formula over $\vec{V} \cup \vec{U}$. For any variable V_i, we denote by ϕ_{V_i} the formula in the equation for V_i in M. We follow the customary practice of leaving the equations for variables that depend directly on the exogenous variables implicit, and simply state the value they take in each particular story.

For an assignment $(\vec{v}, \vec{u})$ of values to the variables in $\vec{V} \cup \vec{U}$, we denote by $\phi^{(\vec{v},\vec{u})}$ the truth value obtained by filling in the truth values $(\vec{v}, \vec{u})$ in the formula ϕ. An assignment $(\vec{v}, \vec{u})$ *respects* M, if for each endogenous variable V_i, its value $v_i = \phi_{V_i}^{(\vec{v},\vec{u})}$. As usual, we only consider models M in which the equations are acyclic, which implies that for each assignment $\vec{u}$ to $\vec{U}$, there is exactly one assignment $(\vec{v}, \vec{u})$ that respects M. Therefore, we refer to $\vec{U} = \vec{u}$ as a *context*. For every value $\vec{u}$ of $\vec{U}$, we call the pair $(M, \vec{u})$ a *causal setting*. We write $(M, \vec{u}) \models \phi$ if $\phi^{(\vec{v},\vec{u})} = \mathbf{true}$ for the unique assignment $(\vec{v}, \vec{u})$ that respects M.

A *literal* L is a formula of the form $V_i = v_i$ or $U_i = u_i$. Our restriction to Boolean variables is made concrete here: the only values v_i we consider are $\mathbf{true}$ and $\mathbf{false}$. Hence our definitions and results can be generalised by simply lifting this restriction. (See the Appendix for some more details.)

We will use the atom V_i as a shorthand for $V_i = \mathbf{true}$, and the negated atom $\neg V_i$ as a shorthand for $V_i = \mathbf{false}$. If V_i is endogenous, we write ϕ_{L_i} for ϕ_{V_i} in both cases.

A causal model M is a tool to represent *counterfactual* relations between variables, in the sense that changing the values of the variables on the right-side of an equation can change the value of the variable on the left-side, but not vice versa. This makes them suitable devices to model *interventions* on an actual setting, meaning changes to the value of a variable V_i that affect only the values of variables that depend on V_i, but not those on whom V_i itself depends.

Syntactically, we make use of the $do()$-operator introduced by Pearl (2000) to represent such an intervention. For a model M and an endogenous variable V_i, we denote by $M_{do(V_i)}$ and $M_{do(\neg V_i)}$ the models that are identical to M except that the equations for V_i are $V_i := \mathbf{true}$ and $V_i := \mathbf{false}$, respectively. Hence for a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C$, the causal setting $(M_{do(\neg C)}, \vec{u})$ corresponds to the counterfactual setting resulting from the intervention on $(M, \vec{u})$ that prevents C.

Throughout this paper, we take C and E to be endogenous literals, where C is a candidate cause for the effect E.

3 Literature Survey

In this paper we consider the following approaches to defining causation: (Lewis, 1973; Hitchcock, 2001; Woodward, 2003; Hall, 2004, 2007; Halpern and Pearl, 2005; Halpern, 2016; Weslake, 2015). All of them take as their starting point the assumption that counterfactual dependence is sufficient for causation. Informally, given that E and C in fact did occur, E is said to be (counterfactually) dependent on

terpreted as the following condition: $R(x, y) \Rightarrow \neg R(y, x)$. Here we take it to mean the following instead: $R(x, y) \Rightarrow \neg R(\neg x, y)$.

C if E would not have occurred without C. The counterfactual is here interpreted in the usual non-backtracking sense, meaning we assume the non-occurrence of C is the result of an intervention $do(\neg C)$ on the actual story. For matters of simplicity, most authors only consider deterministic examples, meaning the intervention $do(\neg C)$ results in precisely one counterfactual story. We comply with this custom for the most part of this paper, but in Section 4.1 we also consider non-deterministic dependence: E is *possibly* counterfactually dependent on C if *possibly* E would not have occurred without C. We here present a formal definition of dependence in the deterministic case.

Definition 1. *Given a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C \wedge E$, E is* counterfactually dependent *on C if $(M_{do(\neg C)}, \vec{u}) \models \neg E$.*

We take the sufficiency of dependence as our first principle.

Principle 1 (Dependence)**.** *If E is dependent on C in a causal setting $(M, \vec{u})$, then C is a cause of E w.r.t. $(M, \vec{u})$.*

There are three basic types of examples which are commonly used to defend the claim that dependence is not necessary, namely *Early Preemption*, *Late Preemption*, and *Symmetric Overdetermination*. Of course there exist many more counterexamples to the necessity of dependence, but in essence they can all be reduced to these paradigmatic cases, or combinations thereof. To illustrate, we present a case of *Early Preemption* from Hitchcock (2001)[p. 276]:

Example 1 (Backup)**.** *An assassin-in-traininig is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim.*

The following is the standard model used in the literature for this story, where the context is such that $Trainee$ is $\mathbf{true}$.

$$Victim := Trainee \vee Supervisor.$$
$$Supervisor := \neg Trainee.$$

Intuitively it is clear that $Trainee$ is a cause of $Victim$, yet using this model we see that $Victim$ is not dependent on $Trainee$. The starting point for any definition of causation in the counterfactual tradition is to provide a way of handling cases of *Early Preemption*. Lewis (1973) does so by invoking another appealing principle of causation: that it is a transitive relation.

Principle 2 (Transitivity)**.** *If C is a cause of D and D is a cause of E w.r.t. $(M, \vec{u})$, then C is a cause of E w.r.t. $(M, \vec{u})$.*

Lewis (1973) takes **Dependence** and **Transitivity** at face value, and defines causation as the transitive closure of dependence. This definition is able to handle cases of *Early Preemption* by focussing on an intermediate event in between Trainee's shot and Victim getting hit, for example the event of Trainee's bullet flying through the air. By adding a variable to the model representing this event, say $Bullet$, Lewis gets the desired result: $Victim$ is dependent on $Bullet$, $Bullet$ is dependent on $Trainee$, and thus by **Transitivity**, $Trainee$ causes $Victim$.

Elegant as it may be, McDermott (1995) demonstrated that there are two major problems with this definition: it is neither a necessary condition for causation, nor a sufficient one. For the former: cases of Late Preemption and Symmetric Overdetermination do not contain a chain of dependencies, and still intuitively exhibit causation. For the latter: there are intuitively convincing counterexamples to the transitivity of causation. The first problem is generally taken to be a decisive blow to Lewis' definition. The second problem, however, has more general repercussions: giving up **Transitivity** is not taken lightly. To understand why this is the case, we give an overview of the literature on this problem.

3.1 Counterexamples to Transitivity

Many authors have tackled the issue of transitivity, and their analysis always contains the following two properties: the transitivity of causation sounds intuitively appealing, but unfortunately there are convincing counterexamples (McDermott, 1995; Hall, 2000, 2004; Hitchcock, 2001; Sartorio, 2005; Halpern and Pearl, 2005; Halpern, 2015; Paul and Hall, 2013). Halpern (2015) is the most recent to take up this view, summarising the importance of transitivity in the counterfactual tradition as follows [p. 2]:

> Paul and Hall (2013)[p. 215] suggest that "preserving transitivity is a basic desideratum for an adequate analysis of causation". Hall (2000) is even more insistent, saying "That causation is, necessarily, a transitive relation on events seems to many a bedrock datum, one of the few indisputable a priori insights we have into the workings of the concept." Lewis (1986, 2000) imposes transitivity in his influential definition of causality, by taking causality to be the transitive closure ("ancestral", in his terminology) of a one-step causal dependence relation.

Although Halpern (2015) agrees that transitivity should be preserved as much as possible, he acknowledges that there are convincing counterexamples, as do all of the other authors mentioned. To illustrate, we present a few of them here.

The first is by Hitchcock (2001), but Hall (2000) gives an

almost identical example.

Example 2 (Boulder)**.** *A boulder is dislodged, and begins rolling ominously toward Hiker. Before it reaches him, Hiker sees the boulder and ducks. The boulder sails harmlessly over his head with nary a centimeter to spare. Hiker survives his ordeal.*

The following is an appropriate model for this story, where the context is such that $Boulder$ is $\mathbf{true}$.

$$Dies := Boulder \wedge \neg Duck.$$
$$Duck := Boulder.$$

We see that Hiker surviving ($\neg Dies$) is dependent on $Duck$, and $Duck$ in turn is dependent on $Boulder$. Hence by **Dependence**, $Boulder$ causes $Duck$ and $Duck$ causes $\neg Dies$. However it would be absurd to conclude from this that the boulder coming down is a cause of hiker's survival. Thus this example presents a violation of **Transitivity**.

The next example is originally due to McDermott (1995)[p. 531], but is also discussed by others (Hall, 2000; Hitchcock, 2001; Halpern, 2015).

Example 3 (Dog Bite)**.** *Terrorist, who is right-handed, must push a detonator button at noon to set off a bomb. Shortly before noon, he is bitten by a dog on his right hand. Unable to use his right hand, he pushes the detonator with his left hand at noon. The bomb duly explodes.*

We model this as follows, where the context is such that $DogBite$ is $\mathbf{true}$.

$$Bomb := LH \vee RH.$$
$$LH := \textit{DogBite}.$$
$$RH := \neg \textit{DogBite}.$$

Just as with Boulder, it would be absurd to consider the dog bite to be a cause of the explosion, as implied by **Dependence** and **Transitivity**.

Next an example from Hall (2000) that is structurally identical to the previous one, and is also discussed by Halpern and Pearl (2005).

Example 4 (Switch)**.** *An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track, instead of the right. Since the tracks reconverge up ahead, the train arrives at its destination all the same.*

The following is an appropriate model for this story, where RT (LT) means that the train goes down the right-hand (left-hand) track, $Dest$ means that the train arrives at its destination, and the context is such that *Switch* holds, i.e., the engineer flips the switch.

$$Dest := LT \vee RT.$$
$$LT := \textit{Switch}.$$
$$RT := \neg \textit{Switch}.$$

Intuitively, flipping the switch is not a cause of the train's arrival, again going against the combined claims of **Dependence** and **Transitivity**.

Many more counterexamples are given in the literature, but their structures are very similar to the examples here presented. Given the existence of these intuitively convincing counterexamples, all of the authors mentioned agree that **Transitivity** should be abandoned.[2]

Although this means we are sacrificing an intuitive property of causation, we should be careful not to sacrifice too much: even if some cases provide convincing counterexamples to transitivity, there is no reason to abandon it altogether. Again we take our cue from Halpern (2015)[p. 2]:

> In light of the examples, should we just give up on these intuitions? Paul and Hall (2013) suggest that "What's needed is a more developed story, according to which the inference from "C causes D" and "D causes E" to "C causes E" is safe provided such-and-such conditions obtain – where these conditions can typically be assumed to obtain, except perhaps in odd cases." The goal of this paper is to provide sufficient conditions for causality to be transitive.

Halpern (2015) only discusses such conditions in case of dependence. By contrast, we provide several necessary and sufficient conditions for dependence to be transitive, and derive from this a sufficient condition for causation to be transitive *in general*.

4 The (In)transitivity of Dependence

By **Dependence** we know already that whenever dependence is transitive, causation will be transitive as well. In all of the papers mentioned, it holds for all of the counterexamples there discussed, that they have one essential thing in common: they are also counterexamples to the transitivity of dependence.[3] This leads to the suggestion that likewise, whenever dependence violates transitivity, so does causation. Taken together this amounts to the following Condition:

[2] Originally Hall did try to hold on to **Transitivity**, by sacrificing **Dependence**. Later, he rejected this view Hall (2000, 2007).

[3] This is in line with Hitchcock (2001)[p. 276], who defines *ordinary* cases of causation as those where the transitivity of dependence is respected.

Condition 1. *If E depends on D and D depends on C w.r.t. $(M, \vec{u})$, then it holds that: C causes E w.r.t. $(M, \vec{u})$ iff E depends on C w.r.t. $(M, \vec{u})$.*

Any definition which satisfies this condition has the desirable property that it violates transitivity in all of the counterexamples discussed in the literature, while also respecting transitivity in ordinary cases where dependence does so as well.

Recall that dependence is not necessary for causation in general, due to problem cases exhibiting *Early Preemption*, *Late Preemption*, or *Symmetric Overdetermination*. The above condition states that in case we have *a chain of dependencies*, dependence does become a necessary condition. To bring these two observations in agreement requires showing that those problem cases do not occur in case there is a chain of dependencies from C to D to E, but no dependence of E on C.

Regarding *Late Preemption* and *Symmetric Overdetermination*, we shall be brief: there is no example in the literature we know of that is considered a case of either of those, and for which Condition 1 is violated. Moreover, *Late Preemption* differs from *Symmetric Overdetermination* only with regards to its temporal properties, which as mentioned earlier is of no relevance to this paper. Therefore we use the following classic example of *Symmetric Overdetermination* to illustrate our point:

Example 5. *[Symmetric Overdetermination] Suzy and Billy both throw a rock at a bottle. Both rocks hit the bottle simultaneously, upon which it shatters. Either rock by itself would have sufficed to shatter the bottle.*

We can model this story using the single equation $BS := \textit{Suzy} \vee Billy$, where BS represents the shattering of the bottle, and *Suzy* (resp. $Billy$) represent Suzy (resp. Billy) throwing a rock. The context is such that both *Suzy* and $Billy$ are $\mathbf{true}$. Intuitively both *Suzy* and $Billy$ are causes of BS, yet BS is not dependent on either of them. It is clear that in this example the failure of dependence has nothing to do with issues of transitivity. Rather, the problem is that there are two completely independent processes which suffice to bring about E, and both of them actually occur, overdetermining E. Adding more detail by inserting variables in between *Suzy* and BS, so that there is a chain of causes leading from *Suzy* to BS, does nothing to change the fact that there will never be a chain of dependencies from *Suzy* to BS.

4.1 Early Preemption

Cases of *Early Preemption* provide the other reason why dependence is not necessary for causation. Therefore we need to show that such cases do not occur when there is a chain of dependencies, and no dependence from the end of the chain on its start, so that Condition 1 can be accepted.

There is one basic causal setting that is considered by most to be the prototypical case of *Early Preemption*: it is the setting used for *Backup* introduced in Section 3. The model was the following, where the context is such that $Trainee$ holds:

$$Victim := Trainee \vee Supervisor.$$
$$Supervisor := \neg Trainee.$$

Since there is no chain of dependencies between $Trainee$ and $Victim$, Condition 1 does not apply and there is no problem with judging $Trainee$ a cause of $Victim$. We will call this the *small model*.

Observe however that this model is quite similar to that used for *Dog Bite* and *Switch*, two of the counterexamples to the transitivity of causation we discussed earlier. The only difference lies in there being an intermediate variable in between the candidate cause and the effect. In fact, as we mentioned in Section 3, the account of Lewis exploits the similarity between these two models to deal with *Early Preemption*: he creates a chain of dependencies from $Trainee$ to $Victim$ by adding an intermediate variable $Bullet$ that represents the bullet flying through midair. Doing so results in what we will refer to as the *large model*:

$$Victim := Bullet \vee Supervisor.$$
$$Bullet := Trainee.$$
$$Supervisor := \neg Trainee.$$

This model is identical to the ones that we have used for *Dog Bite* and *Switch*. Since Condition 1 was formulated precisely to avoid the conclusion that $DogBite$ and $Switch$ are causes, applying it to this model likewise leads to the result that $Trainee$ is not a cause of $Victim$.

We are thus faced with the following problem: the large model is suggested as a model both for examples labelled *Early Preemption*, in which there is no causation, and for examples labelled *Switch*, in which there is causation. (By calling an example a *Switch*, we simply mean that it is an example where intuitively there is no causation, as opposed to *Early Preemption*.) Therefore the only way out is to argue that the large model is not appropriate for either one of these examples.

We present three strategies for arguing that the large model is not appropriate for *Early Preemption*, and one strategy which argues against it being appropriate for *Switch*. Two of the first three strategies present an analysis of *Early Preemption* that does not conflict with Condition 1. The two remaining strategies do conflict with Condition 1, but we will argue that they are problematic, and should therefore be avoided.

First, we start with the strategy defended by Weslake (2015), which is the simplest. According to him, the distinguishing feature of *Early Preemption* is that there is no

intermediate variable between the candidate cause and the effect. Therefore only the small model is appropriate for *Early Preemption*, and only the large model is appropriate for *Switch*, case closed.

Second, we discuss our own preferred strategy to handle this problem. We argue against using the large model for *Early Preemption*, but the argument applies just as well to the small model: we claim that an appropriate model for *Early Preemption* ought to be non-deterministic. Hence contrary to the first strategy, the presence of an intermediate variable between $Trainee$ and $Victim$ is irrelevant to our strategy.[4]

Specifically, we claim that the underlying motivation behind calling $Trainee$ a cause in the *Backup* example is that contrary to the two models above, intuitively we do consider it possible that Victim's death *is dependent* on Trainee's action after all. For instance, it is natural to assume that even Supervisors are not always accurate, or may also have a loss of nerve. Adding these assumptions to the small model gives the following, more appropriate, non-deterministic model:

$$Victim := Trainee \vee (Supervisor \wedge Accurate).$$
$$Supervisor := \neg Trainee \wedge \neg Nerves.$$

Here we have added variables to represent the accuracy of Supervisor's shot, and the possibility that he has a loss of nerve. The actual story only gives us the partial context such that $Trainee$ holds, leaving unspecified the values of $Accurate$ and $Nerves$. Using this model results in there being possible counterfactual stories so that Victim does not die. This implies that possibly $Victim$ is dependent on $Trainee$, and therefore possibly $Trainee$ is a cause of $Victim$.

We can make this formally precise by extending the context $\vec{U}$ with exogenous variables $\vec{W}$ (such as $Accurate$ and $Nerves$) whose values are undetermined in the actual story.

Definition 2. *Given a causal model M over endogenous variables $\vec{V}$ and exogenous variables $\vec{U}$, we define a* partial context *as an assignment $\vec{u'}$ of values to variables so that $\vec{U'} \subseteq \vec{U}$, and refer to $(M, \vec{u'})$ as a* partial causal setting. *We call an assignment $\vec{w}$ to the remaining exogenous variables $W = U \setminus U'$ a* completion *of u'.*

Further, we will say that something is certain *in a partial causal setting $(M, \vec{u'})$, if it holds in all causal settings $(M, \vec{u'} \cup \vec{w})$ that complete u'. Likewise, something is* possible *if it holds in at least one causal setting $(M, \vec{u'} \cup \vec{w})$ that completes u'.*

[4]Menzies (2004) also argues that the intermediate variable does not matter, and likewise claims that $Trainee$ is not a cause in either of the two deterministic models.

Since $Victim$ is dependent on $Trainee$ in case either $\neg Accurate$ or $Nerves$ holds, by **Dependence** we get that $Trainee$ is possibly an actual cause of $Victim$.

One might object that this conclusion is too weak, on the grounds that our intuitive judgment is best understood as the claim that $Trainee$ is *certainly* an actual cause of $Victim$. If that is true, then we bite the bullet and accept the fact that our intuitive judgment is wrong. There is however a more subtle way to understand our intuitive causal judgments that allows them to fall short of certainty, by distinguishing between different levels of information about the actual story. If we have full information, then we are dealing with a complete causal setting, and the notions of certainty and possibility collapse into the actual. In the absence of any information regarding the actual story, on the other hand, we are left with just the causal model to determine the possible actual causes. Therefore the notions of "possible actual cause" and "actual cause" come closer together as a partial context comes closer to a complete context. The formal details that determine when these two notions are close enough to be used interchangeably are to be found in the Appendix, but the idea is, roughly, that we only allow uncertainty regarding exogenous variables whose influence is limited entirely to counterfactual stories. This condition is fulfilled in *Early Preemption* cases like *Backup* because the values of $Accurate$ and $Nerves$ only come into play when considering the counterfactual that Trainee would not have shot.

Another possible objection to our strategy is that it amounts to begging the question, since the elements of uncertainty that we have introduced are not part of the original story. Therefore, the objection goes, our solution no longer works if we explicitly stipulate that Victim will die in all of the possible counterfactual stories. Our answer is simple: we agree that in such a case $Trainee$ would indeed not be a cause of $Victim$.

We believe the problem here lies not with our causal judgments, but with the discrepancy between stipulating that a causal model is deterministic, and our intuition that it is not. By stipulating that certain extremely counterintuitive characteristics hold for a scenario that is greatly underdetermined by its short description, we end up with an example that is too far removed from common-sense for our intuitions to offer any guidance. Concretely, in order to get our intuitions on board with the assumption that the *Backup* example is truly deterministic requires more than stating that it is impossible for Supervisor to have a loss of nerves, or to miss when he shoots. For starters, we need to imagine a situation where the only possible options for Trainee are to either shoot accurately at Victim, or not to shoot at all. We are to imagine that he is unable to shoot and miss on purpose, or that he shoots Supervisor instead of Victim, or shouts to Victim to seek cover, etc. If we are to rely on our intuitions here at all, we better come up with a more real-

istic scenario to describe Trainee's predicament. In other words, we need an example where Trainee is faced with a binary choice such that regardless of the choice he makes Victim will die as a result. We don't have to look very far, the following example fits those criteria:

Example 6 (Trainee's Switch)**.** *Trainee is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track, instead of the right. Since the tracks reconverge up ahead, the train arrives at its destination all the same. Unfortunately for Victim, he is tied to the tracks at the destination, and is killed by the train.*

As was the case for *Switch*, here it is entirely plausible to assume that the backup process will certainly function properly, i.e., the right track will not break down all of a sudden. Hence here we certainly have an example for which our intuitions agree that Victim's death is inevitable, as opposed to the *Backup* example. As with *Switch*, intuitively Trainee flipping the switch is not a cause of Victim's death, which confirms our analysis.

Cases of *Early Preemption* are distinguished from *Switches* on a one-by-one basis: is the scenario such that the relevant counterfactual story, in which we have both $\neg C$ and $\neg E$, should be allowed by the model, i.e., is there any reason to doubt that the backup process will function properly? If yes, then possibly there is dependence and we consider it a case of *Early Preemption*. If no, then the backup process – Supervisor shooting, the functioning of the right hand track, Terrorist using his right hand – is taken to be reliable and it is a Switch. Obviously this distinction involves a certain amount of subjectivity, but given the divergence of intuitions between people regarding the same story we take this to be a benefit of our approach.

Third we consider the strategy by Hall (2007), which is similar to ours, but not quite the same. He also uses the large model only for *Switch*, whereas his model for *Early Preemption* contains an extra variable that serves to turn the backup process on or off. As with our strategy, he agrees that the distinction between the two cases comes down to whether or not the backup process can fail. The difference is that on his view of *Early Preemption*, even if we somehow have evidence that in the actual story the backup process was reliable and *would not have failed*, we may still consider the counterfactual story in which it does fail. But on this view it becomes quite hard – if not impossible – to express the difference between *Early Preemption* and a *Switch*. For every backup process there is some relevant property on which its reliability depends: Supervisor being accurate or not losing his nerves, Terrorist's ability to use his right hand, the right hand track not being broken, etc. All it takes on his account to change a *Switch* into a case of *Early Preemption* is to add a variable representing this property. Then, even if we have evidence that the relevant property is present, we may still consider the counterfactual story in which it is not.[5] Because of this undesirable consequence, we do not find this strategy convincing.

Fourth there is the strategy offered by Hitchcock (2001) and Halpern and Pearl (2005): they argue that both the small and the large model are appropriate for *Early Preemption*, but neither is appropriate to model switching stories such as *Dog Bite* and *Switch*. They are forced to take up this position, because their solution to get $Trainee$ to come out as a cause in the small model applies just as well to the large model. In response to the common practice of using the large model for Switches, they argue case by case why on closer inspection that model is not appropriate for a particular story, or why that story should be considered a case of *Early Preemption* rather than a *Switch*. Let us examine both replies.

Hitchcock (2001) argues against using the large model for *Dog Bite*. However, that argument only applies to accounts that make use of so-called "ENF counterfactuals", which are a particular form of interventions on a structural model that we will not go into.[6] Halpern and Hitchcock (2010)[p. 16] argue against using the large model for *Switch*, on the basis that the variables LT and RT are logically related: "the train cannot be on both tracks at once". First of all, we disagree that the relation between these variables is logical: it is matter of physics, not logic, that a train can only occupy a single track at any given moment. Second of all, this argument does not apply to *Dog Bite*, as one can push a detonator using two hands. In light of this, and in absence of a general argument as to why such models should never be used to model switches, this reply is not convincing.

Regarding the same *Switch* example, Halpern and Pearl (2005)[p. 27] claim that the large model can be appropriate, but only if we consider the possibility that the right hand track will fail as relevant.

> It is this possibility [that of a malfunctioning track] that should enter our mind whenever we decide to designate each track as a separate mechanism (i.e., equation) in the model and, keeping this contingency in mind, it should not be too odd to name the switch position a cause of the train arrival (or non-arrival).

Pearl (2000) makes the same claim regarding a Switch made up of two lamps. The motivation behind this strategy and ours is the same: if we take the failure of the backup process to be a relevant possibility, then we should consider the counterfactual story in which it does. The difference is that we do not seek recourse in structural contingencies (such as ENF counterfactuals) to represent such

[5] Hitchcock (2009)[p. 398] offers a similar criticism.

[6] See Paul and Hall (2013)[ch. 5] for a detailed discussion of the problems these ENF counterfactuals pose to dealing properly with the counterexamples to **Transitivity**.

counterfactual stories, but use a partial context to allow for non-deterministic models. It is important to point out that this agreement is limited to the distinction between *Switches* and *Early Preemption*, and should not be generalised. Halpern and Pearl (2005) use structural contingencies to consider vastly different counterfactual stories as well, which have nothing to do with the issue at hand.

We have now discussed four of the most important strategies to handling *Early Preemption*, and presented a number of arguments against adopting the third or fourth strategy. What matters for the subject of this paper, however, is that the first and second strategies are both viable options for handling *Early Preemption* properly without running into conflict with Condition 1.

To sum up, because of the fact that the counterexamples to **Transitivity** are without exception also counterexamples to the transitivity of dependence, and in light of the lack of opposition from problem cases like *Overdetermination* and *Early Preemption*, we claim that Condition 1 should be accepted.

5 Transitivity in General

5.1 Contributing

A proper understanding of the intransitivity of causation requires looking further than dependence. Dependence stands at one end of a spectrum, as a strong but intransitive relation that is sufficient for causation. At the other end there is the concept of *contributing*, which is a weak and transitive relation. We introduce some new concepts in order to define it, and present it as a necessary condition for causation.

Definition 3. *We define that a consistent set of literals $\vec{L}$ is* sufficient *for a literal L_i w.r.t. M if $\bigwedge \vec{L} \Rightarrow \phi_{L_i}$ and L_i is positive, or $\bigwedge \vec{L} \Rightarrow \neg\phi_{L_i}$ and L_i is negative. Here, $\bigwedge \vec{L}$ denotes the conjunction of all elements of $\vec{L}$.*

For example, in our rock-throwing model for *Symmetric Overdetermination*, $\{Suzy\}$ is sufficient for BS because $Suzy \Rightarrow Suzy \vee Billy$ is a logically valid implication, similarly $\{\neg Suzy, \neg Billy\}$ is sufficient for $\neg BS$ because $\neg Suzy \wedge \neg Billy \Rightarrow \neg(Suzy \vee Billy)$ is trivially valid.

A sufficient set as a whole clearly contributes to a literal being true, but its necessary elements are doing all the work.

Definition 4. *Given $(M, \vec{u}) \models C \wedge E$, we define that C is a* direct actual contributing cause *of E if there exists a set of literals $\vec{L}$ containing C, such that $(M, \vec{u}) \models \vec{L}$ and $\vec{L}$ is sufficient for E, but $\vec{L} \setminus \{C\}$ is not. We call $\vec{L}$ a* witness *for C w.r.t. E.*

Note that only literals which appear in the equation for E can ever be direct actual contributing causes. To illustrate, both $Suzy$ and $Billy$ are direct actual contributing causes of BS in *Symmetric Overdetermination*, with witnesses $\{Suzy\}$ and $\{Billy\}$ respectively. More generally, the connection between two literals need not be direct:

Definition 5. *Given $(M, \vec{u}) \models C \wedge E$, we define that C is an* actual contributing cause *of E if there exist literals $C = L_1, \ldots, L_n = E$ so that each L_i is a direct actual contributing cause of L_{i+1}.*

From now on we speak simply of C contributing to E, rather than saying that C is an actual contributing cause of E. Informally, if C does not contribute to E, it plays no role in determining the value of E. Indeed, we leave it to the reader to verify that all the definitions under consideration – mentioned in Section 3 – satisfy the following principle:

Principle 3 (Contributing)**.** *If C is a cause of E in a causal setting $(M, \vec{u})$, then C contributes to E w.r.t. $(M, \vec{u})$.*

Informally, what this principle states is that all actual causes of E are literals that contributed to satisfying/falsifying an equation, which in turn contributed to satisfying/falsifying another equation, etc., which in the end contributed to satisfying/falsifying the equation for E.

The following is an interesting connection between dependence and contributing, that will prove useful for interpreting subsequent results.

Theorem 1. *E depends on C w.r.t. $(M, \vec{u})$ iff C contributes to E w.r.t. $(M, \vec{u})$ and $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.*

Proofs of all Theorems can be found in the Appendix.

5.2 A Sufficient Condition for Transitivity

Condition 1 states that causation and dependence are equally transitive in case we have a chain of dependencies.

The next step is to look at transitivity in case there is a chain of causes simpliciter, but not necessarily a chain of dependencies. More specifically, we want to find a good sufficient condition for the transitivity of causation *in general*. Since Condition 1 suffices to respect all counterexamples to **Transitivity** from the literature, a naive suggestion would be to simply demand that causation is always transitive when there is no chain of dependencies. To understand why this would not work, we show how the counterexamples can easily be modified so that there no longer is a chain of dependencies, yet intuition would still find that causation is intransitive. All we need to do is add a little *Symmetric Overdetermination* into the mix.

Example 7 (Dog Bite with Backup)**.** *Imagine the story of the Terrorist from* Dog Bite*, but with a little twist: there are two detonators that can be pushed, either of which will set off the bomb. To make sure nothing goes wrong, Backup pushes the other detonator at the same moment as Terrorist does.*

We can re-use our old model, except that we add Backup's action.

$$Bomb := LH \vee RH \vee Backup$$
$$LH := \textit{DogBite}.$$
$$RH := \neg \textit{DogBite}.$$

Just as in the original example, **Dependence** implies that $DogBite$ is a cause of LH. However, $Bomb$ is now no longer dependent on LH, so there is no chain of dependencies from $DogBite$ to $Bomb$ and Condition 1 does not apply. Because $Bomb$ is symmetrically overdetermined by both LH and $Backup$, we also have that LH should still be a cause of $Bomb$. Nevertheless $DogBite$ should still not be considered a cause of $Bomb$.

The lesson learned is that the focus should not be on the presence of a chain of dependencies as such, but rather on the conditions that decide whether or not dependence is transitive. Therefore we now present three different characterisations of the transitivity of dependence.

Theorem 2. *If E depends on D and D depends on C w.r.t. $(M, \vec{u})$, then the following statements are all equivalent:*

1. *E depends on C w.r.t. $(M, \vec{u})$.*
2. *$\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.*
3. *$\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.*
4. *$\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$.*

Note that in general, i.e., without the restriction to chains of dependencies, the statements in this theorem are not all equivalent. Rather we have that $1 \Leftrightarrow 2$, and $2 \Rightarrow 3 \Rightarrow 4$, but also $4 \not\Rightarrow 3 \not\Rightarrow 2$.

This theorem shows that to satisfy Condition 1, it suffices to take any of the three last conditions as a sufficient condition for the transitivity of causation. Since **Transitivity** is intuitively appealing, we want to restrict **Transitivity** as little as possible. Given that the last condition from Theorem 2 is clearly weaker than the other three (in general), this naturally leads to the following condition:

Condition 2. *[Sufficient Condition for Transitivity] If C causes D and D causes E w.r.t. $(M, \vec{u})$, then the following holds:*

If $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$ then C causes E w.r.t. $(M, \vec{u})$.

In light of Theorem 2 and **Dependence**, we can interpret Condition 2 informally as stating that a definition of causation ought to be "at least as transitive as dependence", i.e., its sufficiency condition for transitivity should be at least as weak as that for dependence. (Note that this statement can be endorsed without having to accept Condition 1.) Before we follow through on this lead, we present an example to show that violations of Condition 2 lead to counterintuitive results.

5.3 Counterexample

To illustrate what goes wrong if Condition 2 is not accepted, we look at two very similar examples using the definitions from Halpern and Pearl (2005), Woodward (2003), and Weslake (2015), none of which satisfies Condition 2.

Example 8 (Assassin)**.** *Assassin adds Cyanide to Victim's coffee, which is certain to kill a person. Backup adds a Liquid to the coffee that reacts with Coffee to form Arsenic, another lethal substance. Victim drinks his coffee, which now contains lethal doses of both Cyanide and Arsenic, and dies.*

We can model this as follows, where the context is such that $Liquid$ and $Cyanide$ are true:

$$Dies := Arsenic \vee Cyanide.$$
$$Arsenic := Liquid \wedge Coffee.$$

All of the definitions listed in Section 3 – including the three mentioned above – agree that $Liquid$ is a cause of $Dies$.[7]

By **Dependence**, $Liquid$ is a cause of $Arsenic$. Further, as was the case with *Symmetric Overdetermination*, $Arsenic$ is a cause of $Dies$. Given that $\neg Liquid$ cannot possibly contribute to $Dies$, Condition 2 implies that $Liquid$ is a cause of $Dies$, in agreement with the above definitions.

However, if we change the example only slightly, we get a different result. Imagine that instead of the Liquid reacting with Coffee to form Arsenic, it's the reaction between Liquid and Cyanide that forms Arsenic. In other words, we assume the following model:

$$Dies := Arsenic \vee Cyanide.$$
$$Arsenic := Liquid \wedge Cyanide.$$

For this example the three definitions mentioned do violate Condition 2, because they no longer judge $Liquid$ to be a cause of $Dies$. It is hard to see what could possibly justify this change in causal judgment: in both cases Liquid is added to a coffee containing Cyanide, in both cases Liquid reacts with part of that mixture, and in both cases this results in two lethal substances overdetermining Victim's death. Therefore the intuitive result would be that $Liquid$ remains a cause in this example as well, which is guaranteed by accepting Condition 2.

[7] Strictly speaking the latest definition defended by Halpern (2016) is an exception, since it judges $Liquid$ to be *part of cause* rather than a cause proper. However, he also suggests that these terms ought to be used synonymously.

6 Transitivity and Asymmetry

6.1 Asymmetry

Looking back at the model which we identified as a *Switch*, i.e., the second model we considered in Section 4.1, we note that there is a *remarkable symmetry* between the actual story and the counterfactual story that we get when intervening on C: in both cases there is a chain of counterfactual dependence from the candidate cause (C and $\neg C$, respectively) to the effect E.

This offers an appealing explanation for why C should not be considered a cause of E in this case: causes are difference makers, i.e., the value of C should make a difference as to whether or not it causes E. In case of dependence, C trivially makes such a difference, as in that case C determines whether or not E occurs at all. Cases of overdetermination and the like show that making a difference can be more subtle. These observations motivated Sartorio (2005) to propose the following principle:[8]

Principle 4 (Asymmetry)**.** *If C is a cause of E w.r.t. $(M, \vec{u})$, then $\neg C$ is not a cause of E w.r.t. $(M_{do(\neg C)}, \vec{u})$.*

As **Asymmetry** and **Transitivity** focus on entirely different properties of causation, it is no surprise that they conflict with each other:

Theorem 3. Dependence, **Transitivity**, *and* **Asymmetry** *are mutually inconsistent.*

Theorem 3 teaches us that accepting **Asymmetry** provides an explanation for the fact that there are violations of **Transitivity**.[9] In fact, picking up our earlier discussion, **Asymmetry** together with **Contributing** helps to make sense of Condition 2, which we can rephrase informally as:

> If C causes D and D causes E w.r.t. $(M, \vec{u})$, then the following holds:
>
> **Transitivity** should be respected unless this would violate **Asymmetry**.

There are now enough elements on the table to construct a coherent genesis of causation that explains its limited transitivity.

6.2 Putting it all Together

We started our analysis by noting the strong connection between dependence and causation. Specifically, by **Dependence** and **Contributing**, we know that causation lies somewhere in between dependence and contributing. Further, in the overwhelming majority of cases, all three of these concepts behave as a transitive relation. So as a first approximation, we assume causation to be some relation, say $Trans(X, Y)$, which satisfies the following condition:

Condition 3. *1. $Trans(X, Y)$ is transitive.*

2. If $Trans(C, E)$ then C contributes to E w.r.t. $(M, \vec{u})$.

3. If E depends on C then $Trans(C, E)$ w.r.t. $(M, \vec{u})$.

The following generalisation of Theorem 1 offers a useful connection between dependence and such a $Trans(X, Y)$ relation.

Theorem 4. *If $Trans(X, Y)$ satisfies Condition 3, then:*

E depends on C w.r.t. $(M, \vec{u})$ iff $Trans(C, E)$ w.r.t. $(M, \vec{u})$ and $Trans(\neg C, \neg E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.

The following is a direct consequence of this theorem, which in analogy with **Asymmetry** we may call **Anti-Symmetry**.

Corollary 1 (Anti-Symmetry)**.** *If E depends on C w.r.t. $(M, \vec{u})$, then $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.*

Informally, this result tells us that dependence is built up out of any relation $Trans(X, Y)$ that satisfies Condition 3, in conjunction with the constraint that it should be Anti-Symmetrical.

Since for causation we only require **Asymmetry**, the solution is straightforward: causation is built up out of some relation $Trans(X, Y)$ that satisfies Condition 3, in conjunction with **Asymmetry**. Putting all of this together, we get the following tentative characterisation of a good definition of causation:

Condition 4. *There exists a relation $Trans(X, Y)$ such that each of the following holds:*

1. $Trans(X, Y)$ is transitive.

2. C causes E w.r.t. $(M, \vec{u})$ iff $Trans(C, E)$ w.r.t. $(M, \vec{u})$ and $\neg Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.

3. If $Trans(C, E)$ then C contributes to E w.r.t. $(M, \vec{u})$.

4. If E depends on C then $Trans(C, E)$ w.r.t. $(M, \vec{u})$.

Any definition of causation satisfying the first and second part of Condition 4 is a compromise between **Transitivity** and **Asymmetry**: **Transitivity** is sacrificed only to the extent that is required to satisfy **Asymmetry**. Add to this the other two constraints, and we get a definition that has all the properties we have argued for.

Theorem 5. *Any definition of causation satisfying Condition 4 satisfies* **Dependence**, **Asymmetry**, *and* **Contributing**, *and Conditions 1 and 2.*

[8] Weslake (2015) adopts a similar – but not identical – principle.

[9] Sartorio (2005) also uses *Switches* to argue that violations of **Transitivity** are due to **Asymmetry**.

Since the weakest possible choice for $Trans$ is to take *contributing to*, we state here the most straightforward definition of causation which meets all the demands of Condition 4.

Definition 6. *Given $(M, \vec{u}) \models C \wedge E$, we define C to be an* actual cause *of E w.r.t. $(M, \vec{u})$ if C contributes to E w.r.t. $(M, \vec{u})$ and $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$.*

This definition gives the desired result for all of the examples discussed. We leave the details to the reader. (See the Appendix for details regarding *Early Preemption*.) However we consider Definition 6 too simple in order to be an adequate definition of causation in general. In this paper we focussed solely on issues arising from **Transitivity** and **Asymmetry**, but we have completely ignored issues related to the temporal properties of causation. More specifically, said definition is unable to deal with examples exhibiting *Late Preemption*, where an effect is overdetermined by two (or more) processes but only one of them deserves to be called a cause. In order to deal with those cases as well, and thus arrive at a definition of causation that is suited in general, the $Trans$ relation should take into account temporal information as well. We intend to do so in future work, by developing the transitive notion of *Production* as a form of contributing that excludes preempted events. Specifically, we aim to generalise the notion of production as introduced by Hall (2004), who uses it to highlight features of causation that are not found in dependence.

7 Conclusion

Starting from the observation that despite the intuitive appeal of the transitivity of causation there are many convincing counterexamples to accepting it, we have constructed an analysis in order to explain the precise relation between causation and transitivity. By pointing out the connection between violations of the transitivity of dependence, and violations of transitivity in general, we arrived at a characterisation of the transitivity of dependence that suggested a suitable sufficient condition for the transitivity of causation. Adding to this the principle of asymmetry resulted in a detailed genesis of causation, that narrows down the search to a proper definition of causation considerably. Finally, we have suggested a definition which meets all the requirements discussed, but remains incomplete until it is complemented with temporal properties that can handle *Late Preemption*.

8 Appendix

8.1 Generalising to Multi-valued Variables

As mentioned in Section 2, throughout this work we have limited our discussion to models that only include Boolean variables. We now explain how we can generalise our results to also include structural models that contain multi-valued variables.

We defined a structural equation as taking on the form $V_i := \phi$, where ϕ is a propositional formula over $\vec{V} \cup \vec{U}$, and we used V_i, resp. $\neg V_i$, as a shorthand for $V_i = \mathbf{true}$, resp. $V_i = \mathbf{false}$. In this more general setting, a structural equation takes the form $V_i := F_{V_i}(\vec{W})$, where F_{V_i} is a function from a set of variables $\vec{W} \subseteq (\vec{V} \cup \vec{U})$ to the domain of V_i. (The domains of each variable may vary, but they are all assumed finite.) As before, we only consider models M such that the equations are acyclic.

Since the values of variables are no longer limited to $\mathbf{true}$ and $\mathbf{false}$, the literals C and E that were used througout have to be written out explicitly as $C = c$ and $E = e$. Further, a negated atom like $\neg C$, i.e., $\neg(C = c)$, should be replaced with a formula $C = c'$, and the condition that $c' \neq c$.

For example, counterfactual dependence is now defined as:

Definition 7. *Given a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C = c \wedge E = e$, $E = e$ is* counterfactually dependent *on $C = c$ if $\exists c' \neq c, e' \neq e : (M_{do(C=c')}, \vec{u}) \models E = e'$.*

Except for explicitly writing out the literals, Condition 1 remains unchanged.

The generalisation of the definition of sufficiency is straightforward:

Definition 8. *We define that a consistent set of literals $\vec{L} = \vec{l}$ is* sufficient *for a literal $L_i = l_i$ w.r.t. M if $\vec{L} = \vec{l} \Rightarrow F_{L_i}(\vec{W}_i) = l_i$.*

From there onwards, all definitions and theorems can be generalised in the same manner. Proofs of all theorems remain the same. **Asymmetry**, for example, now becomes:

Principle 4 (Asymmetry)**.** *If $C = c$ is a cause of $E = e$ w.r.t. $(M, \vec{u})$, then there exists a value $c' \neq c$ so that $C = c'$ is not a cause of $E = e'$ w.r.t. $(M_{do(C=c')}, \vec{u})$.*

Finally, the generalised version of our definition of actual causation becomes:

Definition 9. *Given $(M, \vec{u}) \models C = c \wedge E = e$, we define $C = c$ to be an* actual cause *of $E = e$ w.r.t. $(M, \vec{u})$ if $C = c$ contributes to $E = e$ w.r.t. $(M, \vec{u})$ and there exists a value $c' \neq c$ so that $C = c'$ does not contribute to $E = e$ w.r.t. $(M_{do(C=c')}, \vec{u})$.*

8.2 Actual Causation in Non-deterministic Models

As we mentioned in Section 4.1 when discussing *Early Preemption*, we accept a limited form of uncertainty regarding the actual story in judgments of actual causation. We now make this formally precise, by specifying the level

of information that is required for the notions of "actual cause" and "possible actual cause" to be used interchangeably.

Definition 10. *We define that a partial causal setting* $(M, \vec{u'})$ *is* actually complete *if for any choice of* C *and* E*, one of the two following conditions holds:*

- C *certainly contributes to* E *w.r.t.* $(M, \vec{u'})$.
- C *certainly does not contribute to* E *w.r.t.* $(M, \vec{u'})$.

Given our definition of causation (Definition 6), an actually complete setting is one where all that is missing to be certain of the actual causes is information regarding the contributors in the counterfactual stories. This is precisely the situation that we encounter in our non-deterministic model for *Backup* from Section 4.1: the context tells us that $Trainee$ certainly contributes to $Victim$, but in the counterfactual story all we can say is that $Supervisor$ possibly contributes to $Victim$.

This gives rise to the following generalisation of actual causation to actually complete settings:

Definition 11. *Given an actually complete causal setting* $(M, \vec{u'})$ *such that certainly* $(M, \vec{u'}) \models C \wedge E$*, we define* C *to be an* actual cause *of* E *w.r.t.* $(M, \vec{u'})$ *if* C *certainly contributes to* E *w.r.t.* $(M, \vec{u'})$ *and* $\neg C$ *possibly does not contribute to* E *w.r.t.* $(M_{do(\neg C)}, \vec{u'})$.

Applying this definition to the non-deterministic model for *Backup* gives the desired result that $Trainee$ is an actual cause of $Victim$.

The most general version of our definition of actual causation is given by the straightforward combination of Definitions 9 and 11:

Definition 12 (Actual Causation)**.** *Given an actually complete causal setting* $(M, \vec{u'})$ *such that certainly* $(M, \vec{u'}) \models C = c \wedge E = e$*, we define* $C = c$ *to be an* actual cause *of* $E = e$ *w.r.t.* $(M, \vec{u'})$ *if* $C = c$ *certainly contributes to* $E = e$ *w.r.t.* $(M, \vec{u'})$ *and there exist a value* $c' \neq c$ *so that* $C = c'$ *possibly does not contribute to* $E = e$ *w.r.t.* $(M_{do(C=c')}, \vec{u'})$.

It is clear that if the actually complete causal setting is non-partial (i.e., the context stipulates the value of each exogenous variable) and all variables are Boolean, then Definition 12 reduces to Definition 6.

8.3 Proofs of Theorems

Theorem 1. E *depends on* C *w.r.t.* $(M, \vec{u})$ *iff* C *contributes to* E *w.r.t.* $(M, \vec{u})$ *and* $\neg C$ *contributes to* $\neg E$ *w.r.t.* $(M_{do(\neg C)}, \vec{u})$.

Proof. The implication from right to left is trivial, so we only prove the implication from left to right. So assume E depends on C w.r.t. $(M, \vec{u})$, or in other words, $(M, \vec{u}) \models C \wedge E$ and $(M_{do(\neg C)}, \vec{u}) \models \neg E$.

We first prove that C contributes to E w.r.t. $(M, \vec{u})$.

Take L^1 to be minimally sufficient for E, i.e., L^1 is sufficient for E, and for any $L_i \in L^1$, $L^1 \setminus \{L_i\}$ is not sufficient for E. (Such a set can be constructed by removing elements from a sufficient set $\vec{L}$ one by one.) By construction, all endogenous literals in L^1 are direct actual contributors to E.

By $L_{(M,\vec{u})}$ we denote all literals L_i such that $(M, \vec{u}) \models L_i$.

Since $\vec{U} = \vec{u} \subset L_{(M_{do(\neg C)},\vec{u})}$, it follows that if $L^1 \setminus \vec{U} = \vec{u} \subseteq L_{(M_{do(\neg C)},\vec{u})}$, then $E \in L_{(M_{do(\neg C)},\vec{u})}$, i.e., $(M_{do(\neg C)}, \vec{u}) \models E$. Therefore there exists at least one endogenous literal $D \in L^1$ such that $D \notin L_{(M_{do(\neg C)},\vec{u})}$. By the previous paragraph, D is a direct actual contributor to E.

If $D = C$, then we are finished with this part of the proof. So assume $D \neq C$. We can apply the exact same reasoning as we did for E, to find a direct actual contributor F to D such that $F \notin L_{(M_{do(\neg C)},\vec{u})}$. Since contributing is transitive, F contributes to E as well. Given that there are only a finite number of endogenous literals, and that M is assumed to be acyclical, continuing this reasoning will eventually end up with finding C as an actual contributing cause of E. Therefore we conclude that C contributes to E w.r.t. $(M, \vec{u})$.

We can apply the exact same reasoning to prove that also $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u})$, which concludes the proof. □

Theorem 2. *If* E *depends on* D *and* D *depends on* C *w.r.t.* $(M, \vec{u})$*, then the following statements are all equivalent:*

1. E *depends on* C *w.r.t.* $(M, \vec{u})$.
2. $\neg E$ *depends on* $\neg C$ *w.r.t.* $(M_{do(\neg C)}, \vec{u})$.
3. $\neg C$ *contributes to* $\neg E$ *w.r.t.* $(M_{do(\neg C)}, \vec{u})$.
4. $\neg C$ *does not contribute to* E *w.r.t.* $(M_{do(\neg C)}, \vec{u})$.

Proof. Assume E depends on D and D depends on C w.r.t. $(M, \vec{u})$. First, note that by Theorem 1 this implies that C contributes to D and D contributes to E w.r.t. $(M, \vec{u})$. Since contributing is transitive by construction, this implies that C contributes to E w.r.t. $(M, \vec{u})$.

We start with assuming that E depends on C w.r.t. $(M, \vec{u})$. It follows directly from the definitions that this is equivalent to $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.

Now assume we know that $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)}, \vec{u})$. Given that we already know that C contributes to E w.r.t. $(M, \vec{u})$, by Theorem 1 we

see that this is equivalent to $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.

Lastly, assume that $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u})$. It follows directly that $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$. Remains the reverse implication. It suffices to show that if $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$, then $(M_{do(\neg C)}, \vec{u}) \models \neg E$.

We proceed by a reductio: assume that $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$ and $(M_{do(\neg C)}, \vec{u}) \models E$.

D depends on C w.r.t. $(M, \vec{u})$, and thus $(M_{do(\neg C)}, \vec{u}) \models \neg D$. Together with the fact that $(M_{do(\neg C)}, \vec{u}) \models E$, this implies that $(M_{do(\neg C, \neg D)}, \vec{u}) \models E$. Also, since E depends on D w.r.t. $(M, \vec{u})$, we have $(M_{do(\neg D)}, \vec{u}) \models \neg E$. Therefore $\neg E$ depends on C w.r.t. $(M_{do(\neg D)}, \vec{u})$. By Theorem 1, this implies that $\neg C$ contributes to E w.r.t. $(M_{do(\neg C, \neg D)}, \vec{u})$, and thus also $\neg C$ contributes to E w.r.t. $(M_{do(\neg C)}, \vec{u})$, which concludes the proof. □

Theorem 3. Dependence, **Transitivity**, *and* **Asymmetry** *are mutually inconsistent.*

Proof. We have a look again at the *Switch* example. In the story such that $Switch$ holds, by **Dependence** $Switch$ is a cause of LT and LT is a cause of $Dest$. By **Transitivity**, this makes $Switch$ a cause of $Dest$. But if we look at the story $do(\neg Switch)$, then we can apply the same reasoning to get that $\neg Switch$ is a cause of $Dest$. This is in violation of **Asymmetry**. □

Theorem 4. *If* $Trans(X, Y)$ *satisfies Condition 3, then:*

E *depends on* C *w.r.t.* $(M, \vec{u})$ *iff* $Trans(C, E)$ *w.r.t.* $(M, \vec{u})$ *and* $Trans(\neg C, \neg E)$ *w.r.t.* $(M_{do(\neg C)}, \vec{u})$.

Proof. We start with the implication from left to right. So assume E depends on C w.r.t. $(M, \vec{u})$, which is equivalent to $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)}, \vec{u})$. Hence by applying 3 to both statements, we get the desired result.

Remains the implication from right to left. So assume $Trans(C, E)$ w.r.t. $(M, \vec{u})$ and $Trans(\neg C, \neg E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$. By 2, this implies that C contributes to E w.r.t. $(M, \vec{u})$ and $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u})$. Applying Theorem 1 gives the result. □

Theorem 5. *Any definition of causation satisfying Condition 4 satisfies* **Dependence**, **Asymmetry**, *and* **Contributing**, *and Conditions 1 and 2.*

Proof. **Dependence**: Assume E depends on C w.r.t. $(M, \vec{u})$. By 2, we need to show that $Trans(C, E)$ w.r.t. $(M, \vec{u})$ and $\neg Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$. The former is a direct consequence of 4, so remains the latter.

Since E does not hold in $(M_{do(\neg C)}, \vec{u})$, we get that $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$. By 3, this implies that $\neg Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$.

Asymmetry and **Contributing** follow immediately from 2 and 3.

Condition 1: Assume E depends on D and D depends on C w.r.t. $(M, \vec{u})$. By 1 and 4 this implies that $Trans(C, E)$ w.r.t. $(M, \vec{u})$. The implication from right to left in the equivalence from Condition 1 follows from **Dependence**. So we need to prove that $\neg Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$ implies that E depends on C w.r.t. $(M, \vec{u})$.

We proceed by a reductio: assume that $\neg Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$ and $(M_{do(\neg C)}, \vec{u}) \models E$.

D depends on C w.r.t. $(M, \vec{u})$, and thus $(M_{do(\neg C)}, \vec{u}) \models \neg D$. Together with the fact that $(M_{do(\neg C)}, \vec{u}) \models E$, this implies that $(M_{do(\neg C, \neg D)}, \vec{u}) \models E$. Also, since E depends on D w.r.t. $(M, \vec{u})$, we have $(M_{do(\neg D)}, \vec{u}) \models \neg E$. Therefore $\neg E$ depends on C w.r.t. $(M_{do(\neg D)}, \vec{u})$. By Theorem 4, this implies that $Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C, \neg D)}, \vec{u})$, and thus also $Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$, which concludes the proof.

Condition 2: Assume C causes D and D causes E w.r.t. $(M, \vec{u})$, and $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$. By 1, we get that $Trans(C, E)$ w.r.t. $(M, \vec{u})$. By 3, we also get that $\neg Trans(\neg C, E)$ w.r.t. $(M_{do(\neg C)}, \vec{u})$, and thus C causes E w.r.t. $(M, \vec{u})$. □

References

Hall N (2000) Causation and the price of transitivity. Journal of Philosophy 97(4):198–222

Hall N (2004) Two concepts of causation. In: Collins J, Hall N, Paul LA (eds) Causation and Counterfactuals, The MIT Press, pp 225–276

Hall N (2007) Structural equations and causation. Philosophical Studies 132(1):109–136

Halpern J (2015) Sufficient conditions for causality to be transitive. Philosophy of Science

Halpern J (2016) Actual Causality. MIT Press

Halpern J, Hitchcock C (2010) Actual causation and the art of modeling. In: Causality, Probability, and Heuristics: A Tribute to Judea Pearl, London: College Publications, pp 383–406

Halpern J, Pearl J (2005) Causes and explanations: A structural-model approach. part I: Causes. The British Journal for the Philosophy of Science 56(4):843–87

Hitchcock C (2001) The intransitivity of causation revealed in equations and graphs. Journal of Philosophy 98:273–299

Hitchcock C (2009) Structural equations and causation: six counterexamples. Philosophical Studies 144:391–401

Lewis D (1973) Causation. Journal of Philosophy 70:113–126

Lewis D (1986) Causation. In: Philosophical Papers II, Oxford University Press, pp 159–213

Lewis D (2000) Causation as influence. Journal of Philosophy 97(4):182–197

McDermott M (1995) Redundant causation. The British Journal for the Philosophy of Science 46(4):523–544

Menzies P (2004) Causal models, token causation, and processes. Philosophy of Science 71(5):820–832

Paul L, Hall N (2013) Causation: a user's guide. Oxford University Press

Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press

Sartorio C (2005) Causes as difference-makers. Philosophical Studies 123:71–96

Weslake B (2015) A partial theory of actual causation. The British Journal for the Philosophy of Science forthcoming

Woodward J (2003) Making Things Happen: A Theory of Causal Explanation. Oxford University Press