A Principled Approach to Defining Actual Causation

Sander Beckers and **Joost Vennekens**
Sander.Beckers@cornell.edu
Joost.Vennekens@cs.kuleuven.be

Abstract

In this paper we present a new proposal for defining actual causation, i.e., the problem of deciding if one event caused another. We do so within the popular counterfactual tradition initiated by Lewis, which is characterised by attributing a fundamental role to counterfactual dependence. Unlike the currently prominent definitions, our approach proceeds from the ground up: we start from basic principles, and construct a definition of causation that satisfies them. We define the concepts of *counterfactual dependence* and *production*, and put forward principles such that dependence is an unnecessary but sufficient condition for causation, whereas production is an insufficient but necessary condition. The resulting definition of causation is a suitable compromise between dependence and production. Every principle is introduced by means of a paradigmatic example of causation. We illustrate some of the benefits of our approach with two examples that have spelled trouble for other accounts. We make all of this formally precise using structural equations, which we extend with a timing over all events.

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? Although progress has been made over the last decade, not a single definition on offer goes uncontested. In this paper we develop a new proposal for defining actual causation. In comparison to the large number of proposals out there, our approach offers the important benefit that it starts from basic principles. Indeed, many existing definitions lack proper foundations. Even when a detailed justification is given, it mostly consists of informal guidelines rather than precise formal conditions. By contrast we aim to make explicit what principles we take to be fundamental to causation, and show their consequences on particular examples. In this manner even those who disagree with the verdicts of our definition are guided to the principles from which they follow.

As a starting point, we delineate the borders of the search space we wish to explore. This implies formulating a sufficient and a necessary condition. The former serves as a lower bound, in the sense that its extension is a subset of all cases of actual causation, whereas the latter forms an upper bound. These conditions thus form the boundaries of a spectrum of concepts that contains actual causation somewhere in between. The task before us is to provide principles which point towards a single concept in this spectrum.

The literature on actual causation abounds in convoluted examples that discredit or confirm definitions of causation. To make matters worse, these definitions themselves often turn out to be quite hard to understand. To avoid these pitfalls we illustrate every principle by a very simple example, and indicate how the intuition behind it can be made formally precise using structural equations. To obtain maximal clarity, all but one of these examples are made up of the same ingredients, namely two protagonists named Billy and Suzy, each holding a rock in their hand, and a bottle that is standing a bit further waiting to be shattered. Hall and Paul (2003) introduced these types of examples, which are now widespread in the literature. Small changes to the details of the scenario suffice to highlight what we take to be the fundamental issues of the debate. Although we view it as a benefit of our approach that it can be developed using the simplest of examples, we also show how it handles two examples that have spelled trouble for other accounts.

Elsewhere (Beckers and Vennekens, 2016a; Vennekens, 2011), we have discussed a graded, probabilistic notion of causation, but in the current paper we restrict ourselves to a binary concept, i.e., we are purely interested in the question whether or not something is a cause. Further, for the most part we limit ourselves to deterministic examples. Also, we set aside interesting recent research regarding the influence that norms and expectations have on our causal judgments (Halpern and Hitchcock, 2015; Hitchcock and Knobe, 2009). We intend to address these issues in relation to the definition here developed in future work.

Lastly, we point out that there is one important problem which (for the most part) we will ignore in the current work: the (in)transitivity of causation. We discuss this issue in-depth in (Beckers and Vennekens, 2016b), where we offer additional support for the definition of causation as developed here, by considering under what conditions causation violates transitivity. Therefore the two papers are complimentary.

In the next section we introduce the formal framework

of structural equations. Section 3 presents dependence as a sufficient condition for causation, followed by a necessary condition in Section 4. Sections 5 and 6 then present production as a necessary but insufficient condition lying in between the previous ones. Section 7 addresses issues arising from non-determinism, and also compares our approach to that of Halpern and Pearl (2005). In Section 8 we refine our conditions by having a more detailed look at dependence, which narrows down the search space to a single option by discussing another example in Section 9. Section 10 interprets the resulting definition as a compromise between the concepts of *counterfactual dependence* and *production*. To conclude, Section 11 discusses two examples which other definitions are unable to handle.

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.

The 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. If an equation takes the form $V_i := U_j = u_j$ for some U_j and u_j, we shall say that V_i is determined directly by $\vec{U}$. We follow the customary practice of leaving the equations for such endogenous variables implicit, and simply state the value V_i takes 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.

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 $V_i = \mathbf{false}$. Regardless of whether $L_i \equiv V_i$ or $L_i \equiv \neg V_i$, we write ϕ_{L_i} to mean ϕ_{V_i}. Hence in the case where $L_i \equiv \neg V_i$, $\neg\phi_{L_i}$ will be a propositional formula that makes L_i true in any assignment that respects M. Further, we denote by $L_{(M,\vec{u})}$ the set of all literals L_i such that $(M, \vec{u}) \models L_i$.

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.

To illustrate, we present a very simple causal model describing a boy named Billy and a girl named Suzy, who occasionally like to get together and throw rocks at a bottle. We need three endogenous variables: BS represents the event that the bottle shatters, *ST* and BT that Suzy, respectively Billy, throw a rock. In this toy example, we assume that either rock is sufficient for the bottle to shatter. Also, we do not model the causes of them throwing, and just take this to be determined by the background conditions. Thus an appropriate causal model M for this example consists of the single equation $BS := ST \vee BT$.

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

3 Counterfactual Dependence

Consider the first of our rock-throwing stories:

Example 1. *Suzy throws a rock at a bottle, while Billy idly stands by with a rock in his hand, having no intention to throw it. Suzy's rock hits the bottle, at which point it shatters.*

Formally, this example is represented by the context such that only Suzy throws her rock, resulting in the assignment $\{ST, \neg BT, BS\}$. Without hesitation everyone would agree that Suzy throwing her rock caused the bottle to shatter. This judgement can be justified by a straightforward counterfactual observation: if Suzy had not thrown her rock, then the bottle would not have shattered. Formally, we use the following definition.

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

In words, E is counterfactually dependent on C if intervening on the value of C, while holding the context fixed,

results in $\neg E$. In the example it is easy to see that indeed BS counterfactually depends on *ST*, but not on $\neg BT$.

This simple example, and the way we treat it, accounts for the majority of our everyday causal attributions. Hence it should come as no surprise that Hume (1748) defined actual causation – causation, for short – as counterfactual dependence – dependence, for short.[1] Following him, dependence is taken by many to be an important intuition underlying causation (Hitchcock, 2001; Hall, 2004, 2007; Halpern and Pearl, 2005; Halpern, 2016; Lewis, 1973; Pearl, 2000; Weslake, 2015; Woodward, 2003). In fact, all of these authors agree, as do we, that dependence is sufficient for causation.[2] Therefore this assumption serves as our starting point.

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. $(M, \vec{u})$.*

4 Contributing Cause

While dependence is sufficient for causation, it is well-known not to be necessary. Indeed, *Symmetric Overdetermination* (SO) and *Preemption* – both *Late* (LP) and *Early* (EP) – are notorious counterexamples. In this section we compare Example 1 to SO, postponing LP and EP until later.

Example 2. *[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.*

In terms of our causal model, this story represents the context in which both Suzy and Billy throw. Intuitively, most people judge each throw to be a cause of the bottle shattering. However it is easy to see that it is dependent on neither (although it is dependent on at least one rock being thrown, i.e., $M_{do(\neg BT, \neg ST)}, \vec{u}) \models \neg BS$). Despite the lack of dependence, there still is a sense in which we can legitimately say that each throw *contributed* to the shattering of the bottle.

To clarify this notion of contributing, let us zoom out for a second and consider the general causal model, rather than this specific story. At the general level, i.e., in absence of any information regarding the context $\vec{U}$, all we can say is that both *ST* and BT could contribute to BS. Formally, we introduce the concept of a *contributing cause* to express this, which is also defined by Weslake (2015) and Woodward (2003).[3] First, we define the following helpful concept.

Definition 2. *We define that a consistent set of literals L is* sufficient *for a literal L_i w.r.t. M if $\bigwedge L \Rightarrow \phi_{L_i}$ and L_i is positive, or $\bigwedge L \Rightarrow \neg\phi_{L_i}$ and L_i is negative. Here, $\bigwedge L$ denotes the conjunction of all elements of L.*

Recall that M consists of a set of equations of the form $V_i := \phi_{V_i}$, where ϕ_{V_i} is a propositional formula. Then, according to Definition 2, L is sufficient for L_i w.r.t. M just in case the conjunction of literals in L logically entails the propositional formula ϕ_{V_i} (when $L_i \equiv V_i$), or the propositional formula $\neg\phi_{V_i}$ (when $L_i \equiv \neg V_i$).

For example, in our rock-throwing model, $\{ST\}$ is sufficient for BS because $ST \Rightarrow ST \vee BT$ is a logically valid implication, and $\{\neg ST, \neg BT\}$ is sufficient for $\neg BS$ because $\neg ST \wedge \neg BT \Rightarrow \neg(ST \vee BT)$ is trivially valid.

Definition 3. *Given M, we define that C is a* direct possible contributing cause *of E if there exists a set of literals L containing C, such that L is sufficient for E, but $L \setminus \{C\}$ is not. We call L a* witness *for C w.r.t. E.*

Note that this definition is context-independent, and that only literals which appear in the equation for E can ever be direct possible contributing causes. To illustrate, both *ST* and BT are direct possible contributing causes of BS, with witnesses $\{ST\}$ and $\{BT\}$ respectively: $\{ST\}$ and $\{BT\}$ are both sufficient for BS, and $\emptyset$ is not (since $\top \not\Rightarrow ST \vee BT$).

More generally, the connection between two literals need not be direct:

Definition 4. *Given M, we define that C is a* possible contributing cause *of E if there exist literals $C = L_1, \ldots, L_n = E$ so that each L_i is a direct possible contributing cause of L_{i+1}.*

Besides the sufficiency of dependence, all authors mentioned earlier also agree on the principle that if C is an actual cause of E, then C has to be a possible contributing cause of E.[4] Indeed, if C is not a possible contributing cause of E, then under no circumstances does it affect the truth of E.

A natural step is to zoom in again, and refine this concept and its corresponding principle to the level of an actual story, by plugging a specific context $\vec{u}$ into the model M.

Definition 5. *Given $(M, \vec{u}) \models C \wedge E$, we define that C is a* direct actual contributing cause *of E if C is a direct possible contributing cause of E with a witness L such that $(M, \vec{u}) \models L$.*

Using this notion allows us to differentiate between the role of BT in the contexts of Example 1 and Example 2. For Example 1, we have that $(M, \vec{u}) \not\models BT$, and hence there is no witness for BT being a direct actual contributing cause of BS. For Example 2, on the other hand, we have that $(M, \vec{u}) \models BT$, and hence $\{BT\}$ is a witness to the fact that BT is a direct actual contributing cause of BS. Again we can generalize this concept by considering a chain of direct contributing causes.

Definition 6. *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. We now formulate our second principle, which provides a necessary condition for actual causation and therefore delineates the upper border of our spectrum.

[1] Surprisingly in the same breath he formulated a different definition as well, known as the regularity account, which is also still influential.

[2] Halpern (2016) discusses this for all of the HP-approaches, and Weslake (2015) does so regarding most of the others.

[3] Our formulation and the ensuing principle are not entirely identical to theirs, but the difference is negligible.

[4] For details regarding most of the approaches, again see (Weslake, 2015).

Principle 2 (**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 a formula ϕ_{V_i} for some variable V_i, which in turn contributed to satisfying/falsifying another formula ϕ_{W_i}, etc., which in the end contributed to satisfying/falsifying ϕ_E.

The only difference between this principle and the one mentioned after definition 4, is that we have filled in an actual context. Weslake's definition (2015) has this principle directly built into it, as his (STRAND) condition. The reader may verify that all the other definitions mentioned above also satisfy the principle proposed here, as long as one takes into account the restriction to Boolean endogenous variables made earlier.

Although this restriction is of no importance for the overwhelming majority of cases, there is one exception. In models that represent "trumping causation" by means of a three-valued variable, some of these definitions do call an event a cause even though it fails to contribute to the effect. However, the majority of authors agree that this is the wrong answer.[5] Hence if anything, this speaks in favour of **Contributing**.

Applying this principle allows us to exclude certain literals that clearly are not causes, such as $\neg BT$ in Example 1. We thus now distinguish three relations between C and E:

- E is dependent on C. (*ST* in Example 1)
- C does not contribute to E. ($\neg BT$ in Example 1)
- E is not dependent on C, but C does contribute to E. (*ST* and BT in Example 2)

By **Dependence** and **Contributing** we know that there is causation in the first, and not in the second, of these cases. (That the cases are mutually exclusive, and thus the conjunction of our principles consistent, follows from Theorem 1 in Section 5.) Since *ST* and BT are causes in Example 2, we might hope that besides being necessary, contributing is also sufficient for causation. In the following two sections we present two counterexamples to the sufficiency of contributing, and develop two principles which explain what may prevent contributing literals from being causes.

5 Production

The following story is paradigmatic in the literature for what has come to be known as *Late Preemption* (LP).

Example 3. *[Late Preemption] Suzy and Billy both throw a rock at a bottle. Suzy's rock gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw.*

In this story, the process of Billy throwing a rock and shattering the bottle is *preempted* by the process involving Suzy, and this happens *after* the effect has occurred, i.e., after the bottle has shattered. This is in contrast to *Early Preemption* (EP), in which a process is preempted before the effect occurs.[6] (See Section 7.)

As in SO, the bottle shattering is overdetermined by both throws, and again the bottle's state is not dependent on either throw. The difference here is the asymmetry that Suzy's rock hits the bottle, while Billy's does not. Our causal judgments reflect this asymmetry, as people unanimously judge Suzy's throw to be the sole cause.

How should we formally represent this example? If we continue using the three-variable causal model $BS := ST \vee BT$, then we end up with the same causal setting as in SO. Since BT is a cause in SO, but not in LP, we need to refine our representation to take into account the difference between them. More specifically, we need to represent precisely that difference which justifies the shift in causal status of BT when going from SO to LP.

As noted, the difference between SO and LP is whether or not Billy's rock hits the bottle. Hence, one might distinguish between the two cases by adding variables *SH* and BH to the model, which represent Suzy's, respectively Billy's, rock hitting the bottle. Using such a model allows one to make the following observation: if we hold fixed BH at its actual value, then BS is dependent on *ST* in case of LP, but not so in case of SO. This approach is taken by Halpern and Pearl (2005) – HP – whose work on actual causation has set the benchmark for others to compare with. Their definition – in its many versions – takes full advantage of holding fixed variables at specific values regardless of their equations, given that certain structural criteria are fulfilled. They refer to these non-standard causal settings as *structural contingencies*.

We discuss the HP approach in Section 7.1, for now suffice it to say that we believe this approach is flawed, for it does not take into account the *reason why* Billy's rock did not hit the bottle, despite him throwing it. Yet that reason is obvious: Billy's rock arrived at the bottle *too late*. Or, in the words of Hall (2004)[p. 8]:

> Once the bottle has shattered, however, it cannot do so again; thus the shattering of the bottle prevents the process initiated by Billy's throw from itself resulting in a shattering.

If there is any principle regarding causation which is accepted across the board, then it is the fact that causes come before – or at most simultaneous with – effects. Our approach to handle LP consists in combining said principle with the temporal information regarding Billy's rock.

In order to formally represent this idea, it is necessary to represent the completion of each of the competing processes. The most obvious way to do so is by adding one variable for each process: *SA* represents Suzy's rock arriving at the bottle's location, and analogously for BA and Billy's rock. Our new causal model consists of the equations $BS := SA \vee BA$, $SA := ST$, and $BA := BT$.

As with the original model, *ST* and BT are possible contributing causes of BS. All we have done by adding the intermediate variables *SA* and BA is make explicit that the

[5] See (Weslake, 2015)[p.17] for a discussion.

[6] These examples and this manner of distinguishing between them are due to Lewis (1986).

contributions of both *ST* and BT to BS are mediated entirely through *SA* and BA, i.e., a thrown rock can only cause a bottle to shatter by flying through the bottle's location with sufficient momentum. Hence the question as to why BT is not a cause of BS in LP is shifted to the same question regarding BA. The answer follows from a straightforward application of the accepted principle that causes come before effects, since *the bottle had already shattered by the time Billy's rock arrived*.

We will say of prevented processes and the associated literals, like BT and BA, that they have been *preempted* for the effect. Literals that represent a process which completed succesfully, like *ST* and *SA*, will be referred to as *producers* of the effect.

Given the essential role of temporal information, we choose to represent it separately from the variables. In this manner our approach is not dependent on there being suitable variables that capture the consequences of temporal asymmetry, like the variables *SH* and BH mentioned above. This representational clarity proves useful when dealing with cases of late preemption involving an omission, where such variables are unavailable and other approaches fail, as in Example 10 further on.

Definition 7 (Timing)**.** *A* timing τ *for a causal setting* $(M, \vec{u})$ *is a function from* $L_{(M,\vec{u})}$ *to* $\mathbb{N}$.

Informally, a timing can be interpreted as follows. An atom, like BT, represents the fact that some event occurs in our story. Hence, if L_i is an atom, $\tau(L_i)$ simply represents the moment at which the event L_i happens, e.g., the moment that Billy throws his rock. If, on the other hand, L_i is a negated atom, like $\neg BT$, then it represents an omission, i.e., it represents that some event does not occur. Since there is little sense in asking *when* an event does not occur, we take the pragmatic view that in this case $\tau(L_i)$ represents the moment at which the last event occurred that prevents $\neg L_i$ from happening, in the sense that the outcome of this event – together with the outcomes of all previous events – falsifies the formula ϕ_{L_i}.[7] Hence, the timing of omissions is derived from that of events.

We want to point out that aside from this temporal difference, we treat negated atoms and atoms symmetrically throughout this paper, although some authors object to such a view.[8] This issue will pop up further on in the discussion of Example 5.

Also, by always interpreting atoms as events and negated atoms as omissions, the temporal asymmetry here introduced can be viewed as an implicit distinction between a *default* and a *deviant* value of a variable: only variables taking on their deviant value **true** have an independent timing, whereas the timing of variables remaining in their default value **false** is determined by the timing of the former. This perspective proves helpful when considering Example 10 further on. (We point out though that our version of the default/deviant distinction is rather minimal in comparison to other versions, such as for example that of Hitchcock (2007).)

If $\tau(L_i) < \tau(L_j)$, then this means that L_i happened/was prevented before L_j in the actual story. If $\tau(L_i) = \tau(L_j)$, then this means that both happened simultaneously, at least in the sense that the granularity of the story does not allow us to say which happened first.

Because not every story provides – or requires – complete temporal information, we also introduce the following concept.

Definition 8 (Partial timing)**.** *A* partial timing τ *for a causal setting* $(M, \vec{u})$ *is a partial function from* $L_{(M,\vec{u})}$ *to* $\mathbb{N}$.

Now that we have extended a story $(M, \vec{u})$ to include a timing, we can do the same for a counterfactual story $(M_{do(\neg C)}, \vec{u})$: before $\neg C$, everything remains as it was in the actual story, after $\neg C$ the timing remains open.

Definition 9. *Given* $(M, \vec{u}, \tau) \models C$*, we define* $\tau_{do(\neg C)}$ *as the partial timing that is identical to* τ *up until* $\tau(C) - 1$*, has* $\tau_{do(\neg C)}(\neg C) = \tau(C)$*, and is not defined elsewhere.*

Because the structural equations represent causal relationships and causes must always precede their effects, the structure of the equations imposes restrictions on the timings that are possible. In particular, whenever V_i/$\neg V_i$ was caused at some time t, the causes that enabled/disabled ϕ_{V_i} must have already been present at this time. Further, as mentioned, an omission is caused at the same time as the last event which enabled it.

Definition 10. *Given* $(M, \vec{u}, \tau)$*, for every* n*, we denote by* $L^n_{(M,\vec{u})}$ *the set* $\{L_i \in L_{(M,\vec{u})} \mid \tau(L_i) \leq n\}$*. For each endogenous variable* V_i *and the literal* L_i *containing* V_i *such that* $(M, \vec{u}) \models L_i$*, we define* τ *to be* valid *for* V_i *if*

- $L_i = V_i$ *and* $\tau(L_i) \geq \min_{k \in \mathbb{N}}\{L^k_{(M,\vec{u})}$ *is sufficient for* $L_i\}$*; or*
- $L_i = \neg V_i$ *and* $\tau(L_i) = \min_{k \in \mathbb{N}}\{L^k_{(M,\vec{u})}$ *is sufficient for* $L_i\}$.

A timing is valid for $(M, \vec{u})$ *if it is valid for all variables.*

For example, in our rock-throwing story where both Billy and Suzy throw, we require that $\tau(BS) \geq \tau(SA) \vee \tau(BS) \geq \tau(BA)$. In case Billy does not throw, we require that $\tau(\neg BA) = \tau(\neg BT) = \tau(U_i)$, where U_i represents the exogenous event which prevents Billy from throwing. We can generalize the idea of validity to include partial timings.

Definition 11. *A partial timing* τ *is* possible *w.r.t.* $(M, \vec{u})$ *if there exists a timing* τ' *that extends* τ *(i.e.,* $\tau'(L_i) = \tau(L_i)$ *whenever* $\tau(L_i)$ *is defined) such that* τ' *is valid w.r.t.* $(M, \vec{u})$.

Using the timing, we can formalize the notion of production.

Definition 12. *Given* $(M, \vec{u}, \tau) \models C \wedge E$ *with* τ *a valid timing for* $(M, \vec{u})$*, we define* C *to be a* direct producer *of* E *if* C *is a direct actual contributing cause of* E *w.r.t.* $(M, \vec{u})$*, with a witness* L *such that for each* $L_i \in L$*,* $\tau(L_i) \leq \tau(E)$.

More generally we define production in terms of a chain of direct producers.

[7] Here we are using the informal term "prevent" to get across the general idea. The precise interpretation of a timing is given in Definition 10.

[8] Halpern and Hitchcock (2015) provide some of the different views regarding this matter.

Definition 13. *Given $(M, \vec{u}, \tau)$ with τ a valid timing for $(M, \vec{u})$, we define C to be a* producer *of E if there exist literals $C = L_1, \ldots, L_n = E$ so that each L_i is a direct producer of L_{i+1}. For a partial timing τ', we define that C is a* producer *of E w.r.t. $(M, \vec{u}, \tau')$ if there exists at least one valid timing τ that extends τ' such that C is a producer of E w.r.t. $(M, \vec{u}, \tau)$.*

5.1 Comparison to Hall's Production

Our definition of production is inspired by the concept with the same name from Hall (2004). His definition, however, is restricted to positive literals only, i.e., he only considers chains of direct producers $C = L_1, \ldots, L_n = E$ in which all literals L_i are positive. Our definition includes all cases of production covered by Hall's original version, but also allows the literals in a chain to be negative. For example, our definition also applies to cases of so-called *Double Prevention*, which are typically considered to show how dependence and production diverge. (Examples 4 and 10 further on are illustrations.)

As will become clear later, our more tolerant notion of production paves the way to a natural compromise between dependence and production into a single concept of causation. It was the failure to find such a compromise that originally motivated Hall to accept the existence of "Two concepts of causation", a view which he later abandoned (Hall, 2007).

Hall identified a problem with his definition of production, namely that it is context-sensitive. He illustrates this with the following example (Hall, 2004)[p. 31].

Example 4. *First imagine a scenario where we have $E := C \wedge D$, and both C and D are true. Then we zoom in on the details, and learn that the situation also involves an intermediate variable B, such that: $E := C \wedge \neg B$, and $B := C \wedge \neg D$.*

In both versions, E is dependent on both C and D, so according to our definition they are both causes of E, and thus also producers. According to Hall's definition of production, D is a producer of E in the first version only. Yet all the second version does is to make explicit some details that before were left implicit. In terms of the three original variables, the two models behave identically, namely E holds only if both of C and D do. In the second version, D prevents B, which would have prevented E, making it a case of *Double Prevention*. Because the chain from D to E contains an omission, it cannot fall under Hall's definition of production. From this he concludes that production must be context-sensitive, i.e., it depends on the level of detail that we use. Our definition of production, on the other hand, applies equally to both versions of the example. It therefore avoids Hall's relativistic conclusion.

5.2 Preempted Contributors

Producers are literals whose contribution helped bring about the effect. The following definition on the other hand generalizes the failure of Billy's contribution to do so.

Definition 14. *Given $(M, \vec{u}, \tau) \models C \wedge E$, we define C to be* preempted *for E if C contributes to E w.r.t. $(M, \vec{u})$ and it is not a producer of E w.r.t. $(M, \vec{u}, \tau)$.*

The difference between the role of Billy's throw in SO compared to LP, can now be expressed by saying that it changes from being a producer to being preempted. Concretely, any appropriate timing τ for LP will have $\tau(BA) > \tau(BS)$, whereas for SO, $\tau(BA) = \tau(SA) \leq \tau(BS)$. This allows us to exclude BT from being a cause of BS in LP, by applying the formal counterpart of the aforementioned principle.

Principle 3 (Preemption). *If C is a cause of E w.r.t. $(M, \vec{u}, \tau)$, then C is not preempted for E w.r.t. $(M, \vec{u}, \tau)$.*

Combining **Contributing** and **Preemption** results in a stronger necessary condition for causation:

Corollary 1 (Producing). *If C is a cause of E w.r.t. $(M, \vec{u}, \tau)$, then C is a producer of E w.r.t. $(M, \vec{u}, \tau)$.*

Extending the language of structural equations with explicit timings forms a substantial departure from existing structural equations approaches. However, one should not overestimate the role of a timing either. Looking at Principle 3 and Definition 14, we learn that the influence of a timing is limited to the timing of preempted events. Hence in practice it suffices to just give a partial timing over the literals that represent competing processes and their effect, such as BA, *SA* and BS in case of LP.

In all of the examples we have seen so far, producers were always causes. The next section shows that this is not necessarily the case.

6 Switches

Examples involving a switch make up another popular category to gauge intuitions on causation. The following example is paradigmatic (Hall, 2007)[p. 118]:

Example 5 (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 most people agree that flipping the switch is not a cause for the train's arrival. But obviously it is a cause of the train going down the left-hand track, and this in turn is a cause of the train's arrival. Hence this is a counterexample to the transitivity of causation. Given that production is, by definition, transitive, it is also a counterexample to the sufficiency of production.

Part of the reason why we judge there to be no causation here is that the train would have arrived at its destination either way, i.e., there is no dependence. However we already know that dependence is not necessary for causation, so this is not the whole story. The further justification for our judgment is that the actual and the counterfactual story are too symmetrical in regards to the function of the switch. Flipping the switch contributes to a process that results in the train arriving. Not flipping the switch contributes to a different process, but one that has the exact same result. Therefore *Switch* and $\neg$*Switch* perform the same *causal role* in both stories, that of contributing to a process which results in $Dest$.

A fundamental property of causation, which underlies Principle 1 as well, is that causes are *difference makers*. Dependence expresses the strongest form of making a difference: to make a difference as to whether or not the effect takes place. What the switch example illustrates is that there is a weaker form of making a difference that is a necessary condition for causation, namely that the absence of a cause fulfills a different role than the cause itself. We formalize this property by means of the following principle.

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

This principle and the importance of difference making is defended as well by Sartorio (2005). Also Weslake (2015) incorporates a very similar principle into his definition of causation. However his formulation falls prey to a counterexample that we will discuss in Section 11.

Given the extreme symmetry between the actual story in Example 5 and the counterfactual story where the switch is not flipped, most definitions of causation will either judge both *Switch* and $\neg$*Switch* to be causes in their respective stories, or neither. Accepting Principle 4 delivers the intuitive verdict that neither is a cause. Note though that both *Switch* and $\neg$*Switch* are producers in their respective stories.

The qualification "most definitions" we made in the previous paragraph arises from the fact that some authors claim only events can be causes, and therefore Principle 4 would be trivially satisfied for them. Recall that contrary to this view, we assume **true** and **false** are to be treated symmetrically.

Before we move on, we need to address a possible objection. Some readers may not share our intuitions on the Switch example, on grounds that it is not at all certain the train will arrive either way. For instance, who is to say the right track would not break down? This is an important point, which can be made more vividly by using another famous counterexample to the necessity of dependence, so-called early preemption. We direct our attention to this example, in order to show that we can do justice to these intuitions without dropping **Asymmetry**.

7 Non-determinism

Imagine yet another variant of our story:

Example 6. *Suzy throws a rock at a bottle. The rock hits it, and the bottle breaks. However Billy was watching Suzy, and would have thrown a rock just in case Suzy did not throw.*

This is an example of *Early Preemption*, because the causal mechanism connecting Billy's throw to the bottle shattering is preempted by Suzy already before the effect of the bottle shattering occurs. We can re-use the model from Example 3, except that the equation for BT becomes $BT := \neg ST$.

Only *ST* is directly dependent on the context $\vec{u}$, which is such that Suzy throws. Most authors consider examples of early preemption on a par with late preemption, and hence judge *ST* to be a cause of BS in this case as well. Yet if we compare this example to Switch, then we see that they are remarkably similar. *ST* plays exactly the same role here as *Switch* does, namely it determines which of two processes occurs, where each process by itself is sufficient for the effect to take place.

Everything we just said about Switch, also holds for EP: in both cases the candidate cause – *ST* or *Switch* – is a producer of the effect (just as with SO and LP), and in the counterfactual situation the negation of the candidate cause is also a producer of the effect (contrary to SO and LP). In fact, the structural models and assignments to variables are almost completely identical in both cases.

But then how do we explain the prevalent intuition that *ST is* a cause of BS in EP? Here it becomes useful to consider the possibility that the effect will not occur either way. According to our current model, it is certain that if Suzy does not throw then the bottle will shatter nonetheless. Surely that does not sound very realistic, as who is to say that Billy will not miss? All of our rock-throwing equations so far have assumed that Suzy or Billy throwing always results in the bottle's shattering. This assumption was a harmless simplification in the previous examples, because in each of them the *actual* story contained information on Billy's accuracy (with the exception of Example 1, where it was irrelevant). Because this is no longer the case here (since in the actual story Billy did not even throw), a proper analysis of EP requires incorporating this uncertainty. Hence we extend our model with variables *SAcc* and $BAcc$, representing Suzy and Billy's accuracy when they throw.

$$
\begin{aligned}
BS &:= SA \vee BA.\\
SA &:= ST \wedge SAcc.\\
BA &:= BT \wedge BAcc.\\
BT &:= \neg ST.
\end{aligned}
$$

Allowing for the throws to be inaccurate changes the example significantly. In this paper we have limited ourselves to deterministic examples, meaning we assumed that for each variable there was a definite truth-value in the actual story. The underlying motivation for this limitation is that as a result there is an unambiguous interpretation of the counterfactual story $(M_{do(\neg C)}, \vec{u})$, because that story corresponds to precisely one assignment of truth-values to all variables. To tackle *Early Preemption* we need to take a little excursion into the more general realm of non-deterministic

examples, where there might be several counterfactual stories. Since there is no sense in which $BAcc$ has a value in the actual story where Billy does not throw, the value of $BAcc$ is undetermined. This means we have to consider both the counterfactual story where Billy's rock hits the bottle and it shatters, and that in which he throws and misses.[9]

Our approach can easily be generalised to allow for non-deterministic cases, by extending the context $\vec{U}$ with exogenous variables $\vec{W}$ whose values are undetermined in the actual story (eg., $BAcc$ in EP).

Definition 15. *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'.*

Dependence is then defined as follows:

Definition 16. *Given a partial causal setting $(M, \vec{u'})$ such that for all completions $\vec{w}$ of $\vec{u'}$ we have: $(M, \vec{u'} \cup \vec{w}) \models C \wedge E$, E is* counterfactually dependent *on C if there exists a completion $\vec{w}$ such that: $(M_{do(\neg C)}, \vec{u'} \cup \vec{w}) \models \neg E$.*

All other definitions can be similarly generalised to partial causal settings. As before, actual causation is relative to a *story*. Up until now such stories have been represented formally as a causal setting $(M, \vec{u})$. In the current more general setting, a story takes the form of a partial causal setting extended with a timing: $(M, \vec{u'}, \tau)$.

Our original Principle 4 is then replaced with:

Principle 4 (Asymmetry version 2). *If C is a cause of E w.r.t. $(M, \vec{u'}, \tau)$, then there exists a completion $\vec{w}$ of $\vec{u'}$ so that $\neg C$ is not a cause of E w.r.t. $(M_{do(\neg C)}, \vec{u'} \cup \vec{w}, \tau_{do(\neg C)})$.*

As a consequence, by adding the appropriate variables allowing for several counterfactual stories, we are able to do justice to our intuitions in both *Switch* and EP. If it is realistic to assume that train tracks do not malfunction, then the train will arrive either way and flipping the switch is not a cause. If on the other hand our intuitions do not support this assumption, then possibly the train would not arrive but for flipping the switch, and hence flipping it is a cause.

In the EP example, the counterpart of the malfunctioning track is Billy missing the bottle. Since it is quite plausible to take the accuracy of a boy throwing a rock to be much more uncertain than a sturdy track breaking, it is to be expected that intuitions for Suzy's throw being a cause are more common than those for flipping the switch. Hence an appropriate model for EP should contain a variable representing the uncertainty of the counterfactual story, contrary to a model for *Switch*. The more general non-deterministic version of **Asymmetry** then gives the right answer in both cases.

[9]The value of $BAcc$ being undetermined can either be interpreted ontologically, meaning there is no fact of the matter what its value would have been had Billy thrown, or epistemically, meaning we simply do not possess any information that establishes the value of $BAcc$. Our approach can be applied using either interpretation.

The lesson to be learned here is that structurally there is no difference between examples labelled "switches" and those commonly taken to exhibit early preemption. The difference lies in the reliability of the background process which might produce the effect in the absence of the actual process. Having expounded the importance of non-determinism in these examples, to keep things simple from here onwards we focus again on the deterministic version of **Asymmetry**.

7.1 Comparison to HP

For reasons of simplicity, most structural equations approaches stick to deterministic models. Still, all of them claim to provide an adequate analysis of both *Early Preemption* and *Switch*. To further justify our use of non-determinism, we take a closer look at the most influential account of actual causation, by Halpern and Pearl (2005). They apply the same reasoning to *Switch* as we have applied:

> Is flipping the switch a legitimate cause of the trains arrival? Not in ideal situations, where all mechanisms work as specified. But this is not what causality (and causal modeling) are all about. Causal models earn their value in abnormal circumstances, created by structural contingencies, such as the possibility of a malfunctioning track. It is this possibility 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).

Note that they explicitly refer to "the possibility of a malfunctioning track" as a structural contingency. On the face of it this suggests that the motivation behind their approach for dealing with Early Preemption/Switch is very similar to ours: if it is considered a significant possibility that the backup mechanism fails, then this possibility should be taken into account to assess causation. Concretely, in that case we should take into account the counterfactual story where the candidate cause does not occur, and the backup mechanism fails. Which factors determine whether or not the failure of the backup mechanism – be it a train track or a person throwing a rock – is a significant possibility is mostly an empirical matter, and should be decided on a case by case basis.

We find further confirmation of our interpretation by considering another example of *Early Preemption*, discussed by Halpern and Pearl (2005)[p. 30]. We present here the original formulation by McDermott (1995).

Example 7. *[Early Preemption 2] Suppose I reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window.*

Is catching the ball a cause of the window being safe? Even without giving a structural model to go along with this story, the similarity to *Early Preemption* and *Switch* is obvious. Again, the answer depends on whether or not we consider the possibility that the backup mechanism – the wall blocking the window – will fail. Intuitively, most people judge this example to be more similar to *Switch* than to

Early Preemption, meaning they do not judge catching the ball to be a cause. This is consistent with our approach: as with the failure of train tracks, people generally do not consider it a significant possibility that a solid brick wall will fail to stop a cricket ball. Halpern and Pearl (2005) also treat this example similar to *Switch*:

> If we make both the wall and the fielder endogenous variables, then the fielder's catch is a cause of the window being safe, under the assumption that the fielder not catching the ball and the wall not being there is considered a reasonable scenario. On the other hand, if we take it [sic] for granted the wall's presence (either by making the wall an exogenous variable, not including it in the model, or not allowing situations where it doesn't block the ball if the fielder doesn't catch it), then the fielder's catch is not a cause of the window being safe. It would remain safe no matter what the fielder did, in any structural contingency.

The difference between their approach and ours lies in the method used to represent the failure of the backup mechanism.[10] We choose to do so in a very straightforward fashion: all possible stories are modelled as some partial causal setting $(M, \vec{u}', \tau)$. Hence we interpret the deterministic model for *Early Preemption* as stating that it is *impossible* for Billy to miss when he throws. If this statement is considered inappropriate, then one should use the nondeterministic model given above, i.e., one should add a variable that represents Billy's accuracy and consider the possibility that he misses.

Since Halpern and Pearl restrict themselves to deterministic models, the choice between these two models is not available to them. This explains why they seek recourse in structural contingencies, as they need some other method to consider stories beyond the ones allowed by a structural model.

One could take this to imply that the difference here is merely a matter of taste, depending on one's preferred method to represent uncertainty. This is far from the truth. Halpern and Pearl use structural contingencies in a wide variety of cases, and these go well beyond examples resembling *Early Preemption*.

The criteria for deciding if a structural contingency may be used are rather technical, and are not founded on underlying principles or heuristics that guide their application. As a result, they allow for a plethora of situations for which it is hard to see why we should consider them at all.[11] Indeed, Halpern and Pearl (2005)[p.24] concede that in some cases their definition offers acceptable answers only if one explicitly stipulates which situations are "allowable settings". Therefore the interpretation of structural contingencies we have just given only applies to a limited number of cases.

To illustrate, we briefly return to *Late Preemption*. HP use the following model for this example, where SH and BH represent Suzy's, respectively Billy's, rock hitting the bottle:

$$BS := \mathit{SH} \vee BH.$$
$$SH := \mathit{ST}.$$
$$BH := BT \wedge \neg SH.$$

We first have a look at whether or not this model is appropriate to capture the causal structure behind *Late Preemption*.

A first problem with this model is that the asymmetry between Suzy's throw and that of Billy is built right into the model: it does not allow for the story in which Billy throws faster than Suzy, or the story in which they both throw equally fast, as in *Symmetric Overdetermination*. There is nothing in the informal story in Example 3 to suggest that the difference in speed is a general, structural property. On the contrary, it sounds natural to assume that this difference is a contingent property of the actual story. However, as pointed out by Halpern (2016), this problem can be set straight by also including BH into the equation for SH, and adding an exogenous variable to represent the order by which the rocks arrive. Hence this problem is of little consequence.

A second, more fundamental, problem, is the presence of SH in the equation for BH. Recall that a structural equation represents a causal mechanism, in this case the mechanism connecting Billy's throw to Billy's rock hitting the bottle. That mechanism does not involve SH, since Suzy and her rock form an entirely different and independent mechanism. Therefore, it seems conceptually wrong to include SH in the equation for BH. A consequence of this conceptual error is that if we consider the context where only Billy throws, then $\neg \mathit{ST}$ is actually a producer (according to our definition) of BS. This is a very counterintuitive result.

The role played by $\neg SH$ in the equation for BH is not that of a contributor to BH, but rather that of a constraint: it is supposed to capture the property that a bottle cannot shatter if it has already done so. This confirms that one cannot adequately deal with *Late Preemption* without making vital use of temporal information, as we argued in Section 5. Since HP stick to structural equations proper, they are forced to build this temporal information into the model itself. More specifically, the presence of $\neg SH$ in the equation for BH compensates for the fact that they do not use a timing. Given these counterintuitive consequences, we prefer to use our symmetric model, containing the variables SA and BA.

[10] This difference is not limited to Halpern and Pearl. Collins (2000) and Hitchcock (2001) use the same terminology when discussing which counterfactual scenarios ought to be considered. For example, confronted with Example 7, Collins (2000)[p. 8] says that "It is more far-fetched, on the other hand, to suppose that the brick wall be absent, or that the ball would miraculously pass straight through it." Considering an example involving a boulder Hitchcock (2001)[p. 298] says of the failure of the backup mechanism that "This possibility is just too far-fetched." Hall and Paul (2003)[p. 26] criticise Hitchcock by pointing out the arbitrariness in his use of this terminology.

[11] For details on these situations and the counterexamples they allow, see for example (Hall, 2007; Weslake, 2015). Halpern (2016) has recently proposed a new definition which is more restrictive, avoiding some of these pitfalls, but not all. Further, it allows for new counterexamples, eg., it fails to judge each of *ST* and BT a cause in case of SO.

Here it is useful to point out that for every approach using structural equations, the verdict given by a definition of causation is to a large degree dependent on the particular model being used. Since in many cases there is room for debate as to which model is appropriate for a given informal story, this means one can often counteract undesired outcomes of applying a definition by calling into question the model being used. (See (Halpern and Hitchcock, 2010) for a discussion of this issue.) However because our approach ultimately relies on basic principles, rather than on the intutiveness of examples, we believe it is less affected by this issue. If one accepts our principles, then one can make judgments about a causal model regardless of which informal story it is supposed to capture. In this manner, the problem of model appropriateness can to some extent be separated from the problem of defining actual causation.

Setting aside our disagreement regarding the choice of model for *Late Preemption*, we now turn to the HP approach and how it applies given their preferred model. It considers the structural contingency that Billy throws and yet fails to hit the bottle, even though Suzy does not throw. Contrary to the interpretation used for *Switch*, this structural contingency cannot be interpreted simply as the possibility that the backup mechanism fails to function properly, because the actual story explicitly stipulates that it does not: "Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw." This stipulation is not just a detail occurring in our version of the example, but forms an essential part of *Late Preemption* cases.

One might object that there is also another possible interpretation, consistent with what has been said: namely that a structural contingency represents what is *generally possible*, rather than what is *possible given the actual story*. On this reading, the actual information that Billy was accurate is of no interest, all that matters is whether or not in general it is possible that he is not accurate. But going down this road leads to a slippery slope, for it blurs the distinction between *general* and *actual* causation. More specifically, if one can ignore the actual state of Billy's accuracy, then why not ignore other aspects of the actual story as well? For example, why not then consider the story in which Suzy throws but misses and Billy does not throw, and use it to conclude that Billy's throw also caused the bottle to shatter in *Late Preemption*?

Obviously according to the HP definition it is not the case that anything goes. Only those structural contingencies satisfying certain – somewhat complicated – conditions may be considered. But what should be clear by now, is that it is not easy to come up with a consistent and systematic interpretation of what these structural contingencies are supposed to represent. Therefore we prefer to stay far away from them, and instead simply use a non-deterministic model to represent aspects of the story which are not actually determined, and use a timing to exclude those events which happened too late.

8 Dependence Revisited

So far, we have established that dependence is sufficient for causation but not necessary, while production is necessary but not sufficient. Therefore causation must lie in between these two concepts. To pinpoint its location, we present a theorem that relates dependence to production.[12]

Theorem 1. *Given a valid timing τ, E is dependent on C w.r.t. $(M, \vec{u})$ if and only if both of the following conditions hold:*

- *[Condition 1]: C is a producer of E w.r.t. $(M, \vec{u}, \tau)$.*
- *[Condition 2]: $\neg C$ is a producer of $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u}, \tau_{do(\neg C)})$.*

Because this theorem indiscriminately applies to all valid timings, the first conclusion we can draw from it is that all information contained in a particular timing is lost when we consider dependence. Since we introduced the notion of a timing precisely to distinguish between cases where dependence was too crude a tool, this should not come as a surprise. On the contrary, the lesson learned from comparing examples such as *Symmetric Overdetermination* and *Late Preemption* was that the actual timing should be taken into account in order to judge actual causation. This theorem shows that without loss of generality, we can restrict our attention to one particular timing when comparing dependence and production. We now consider how the conjunction of Conditions 1 and 2 can be weakened, so that we shift from dependence to causation.

By **Producing**, we know Condition 1 should stay. Yet as the *Switch* example has shown, production does not satisfy **Asymmetry**: both $Switch$ and $\neg Switch$ are producers of $Dest$ in their respective stories. Therefore a straightforward and natural suggestion is to combine production (Condition 1) with the constraint that **Asymmetry** should be satisfied. In other words, Condition 2 should be replaced with Condition 2': $\neg C$ is not a producer of E w.r.t. $(M_{do(\neg C)}, \vec{u}, \tau_{do(\neg C)})$.

Since *Switch* was the only example discussed which required us to look beyond production, it is easy to see that defining causation as the conjunction of Conditions 1 and 2' agrees with our judgments on all examples discussed so far. Note however that temporal information plays no role in judging *Switch*: the model is such that each story only allows one valid timing, and hence in this case the notions of producing and contributing are equivalent. Therefore we cannot rule out a slightly stronger alternative to Condition 2, let us call it Condition 2", where producing is replaced with contributing: $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)}, \vec{u})$.

To decide between these two conditions, we now present an example in favour of adopting Condition 2', instead of Condition 2".

9 Not Contributing vs. Not Producing

In this section we present a counterexample to the necessity of Condition 2", resulting in the acceptance of Condition 2'. However the example is rather exotic, since it is hard to even find examples for which these two options disagree. (We have not found any in the literature.) Hence we do not put much weight on our preference of Condition 2' over Condition 2", as in practice this will hardly ever matter.

[12] A proof of this theorem is given the Appendix.

Example 8. *In general, Billy throws rocks at bottles either if Suzy does not, or if he just feels like it. Today, Billy throws a rock at a bottle. Immediately afterwards Suzy throws a rock as well. Suzy's rock was thrown harder, and gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been for Suzy.*

To model this variant of the rock-throwing story, which combines elements of early and late preemption, we need to adjust the equation for BT, giving:

$$BS := \mathit{SA} \vee BA.$$
$$\mathit{SA} := \mathit{ST} \wedge \mathit{SAcc}.$$
$$BA := BT \wedge \mathit{BAcc}.$$
$$BT := Feels \vee \neg \mathit{ST}.$$

Here $Feels$ means that Billy just feels like throwing, regardless of what Suzy does. Hence the context is such that $Feels$ and *ST* hold. An appropriate timing τ is such that $\tau(Feels) \leq \tau(BT) < \tau(\mathit{ST}) \leq \tau(\mathit{SA}) \leq \tau(BS) < \tau(BA)$. The question is whether or not *ST* is a cause of BS.

Given that Suzy's throw preempted Billy's throw from shattering the bottle, the example looks similar to LP, which suggests that *ST* is a cause. On the other hand, in the counterfactual story $do(\neg \mathit{ST})$, Suzy's not throwing contributes to the process that would have Billy's rock shattering the bottle, just as with EP. Even more, we know that Billy was accurate, so there is no counterfactual story in which the bottle does not shatter, contrary to EP. Therefore the example is also similar to a switch, which suggests that *ST* is not a cause.

We believe the first similarity, to LP, to be the more fundamental one: even though it may hold in general that $\neg \mathit{ST}$ can produce BT, in this story we already know that Suzy threw after Billy did. So in this case, Suzy throwing or not throwing was completely irrelevant to Billy's throw, which was instead produced by the fact that he felt like throwing. Therefore when considering what would have happened if Suzy had not thrown, the right answer is that $\neg \mathit{ST}$ would not have produced anything (except for $\neg \mathit{SA}$), and just as with LP *ST* should be judged a cause of BS.

Now we compare how Condition 2" and 2' deal with this example. It is clear that *ST* produced BS in the actual story. In the counterfactual story, $\neg \mathit{ST}$ contributes to BS. Therefore this is a counterexample to the necessity of Condition 2".

The partial timing $\tau_{do(\neg \mathit{ST})}$ has $\tau_{do(\neg \mathit{ST})}(Feels) \leq \tau_{do(\neg \mathit{ST})}(BT) < \tau_{do(\neg \mathit{ST})}(\neg \mathit{ST})$. Therefore $\neg \mathit{ST}$ does not produce BT w.r.t. $(M_{do(\neg \mathit{ST})}, u, \tau_{do(\neg \mathit{ST})})$, which implies it does not produce BS either. This is in agreement with the necessity of Condition 2'. We conclude from this that the right choice to make is to take the conjunction of Conditions 1 and 2' as a sufficient and necessary condition for causation.

10 Discussion and Results

Our principles have led us to propose the following definition of actual causation.

Definition 17. *Given $(M, \vec{u}, \tau) \models C \wedge E$, we define C to be an* actual cause *of E w.r.t. $(M, \vec{u}, \tau)$ if*

C produces E w.r.t. $(M, \vec{u}, \tau)$ and $\neg C$ does not produce E w.r.t. $(M_{do(\neg C)}, \vec{u}, \tau_{do(\neg C)})$.

The precise formulation of this definition is dependent on the fact that we defined production over a partial timing as being a producer in at least one valid timing that extends it (as opposed to being a producer in all of them). This boils down to assuming that the default is for actual contribution to imply production, which is in line with our earlier observation regarding the limited influence of a timing: unless we know that a contributing process was preempted, it is a producer. As with the difference discussed in the previous section however, there are very few examples where this distinction matters.

Our definition of actual causation is built up entirely out of production, a concept which has so far received too little attention in the literature. A key property of production is that it focusses solely on the actual world: unsatisfied literals are entirely irrelevant. It tells us whether some event *brought forth* another as things actually happened.[13]

Causation shares production's interest in the actual world, but extends it with a contrast to a counterfactual world: did some event bring forth another as things actually happened, and if so, would the absence of said event not have brought forth the other? In the overwhelming majority of cases, if the first question is answered in the affirmative, so is the second; only examples exhibiting switching behaviour form an exception. This seems to suggest that the intense focus on the counterfactual nature of causation that we have observed in recent years is somewhat misguided. However when we take into consideration Theorem 1, the picture becomes more nuanced, since dependence and production are tightly connected as well.

The main distinguishing feature of dependence is that it cares only about end results: it considers only whether C and E hold in the actual and counterfactual story, without looking at the temporal details – i.e., the timings – of how this came about. The importance of dependence therefore lies in its simplicity: one can forget about the intricacies of timing and preemption, and still end up with an answer that does the job most of the time.

We can express the difference between production, dependence, and causation in a nutshell by saying that production answers the "How?" question, dependence answers the "What if?" question, and causation answers the "Why?" question. The first is usually associated with understanding, the second concerns a form of a posteriori prediction, and the third is fundamental to explanation.

Although we have built up our definition using formal principles and theoretical examples, there has been empirical validation recently that points in a very similar direction. The idea that causation is a combination of dependence and production has been confirmed experimentally on a set

[13] In this respect it is similar to the notion of responsibility as it figures in ethics: ethical judgments concern (for the most part at least) what did happen, not what could have happened. We intend to examine this similarity in more detail in future work.

of physical test-cases by Gerstenberg et al (2015), although their notion of production is less formal and somewhat different from ours. They too stress the importance of distinguishing between different ways a cause can make a difference to the effect (Gerstenberg et al, 2015)[p. 1]:

> We argue that the core notion that underlies people's causal judgments is that of difference-making. However, there are several ways in which a cause can make a difference to the effect. It can make a difference to *whether* the effect occurred, and it can make a difference to *how* the effect occurred.

We now have a look at our definition in practice by confronting it with some troublesome examples.

11 Some Examples

Weslake (2015) gives an overview of the most prominent definitions of actual causation in the structural equations framework. After presenting counterexamples to all of them, he proposes a definition that succeeds in getting the right answer for these examples. We leave it to the reader to verify that our definition delivers the correct verdict in these cases as well.[14] More interesting are his "non-structural counterexamples". These exhibit structural patterns that are identical to cases of symmetric overdetermination and early preemption, yet seem to give rise to different intuitions. He leaves it as an unsolved problem how to deal with these examples correctly as well.[15] Therefore we consider them as suitable test-cases for our approach.

The first example, named "Careful Poisoning", has the same structure as early preemption (Weslake, 2015)[p. 22].

Example 9. *Assistant Bodyguard puts a harmless antidote in Victim's coffee (A). Buddy then poisons the coffee (B), using a poison that is normally lethal, but which is countered by the antidote. Buddy would not have poisoned the coffee if Assistant had not administered the antidote first. Victim drinks the coffee and survives ($\neg D$).*

Intuitively, most people – but not all – agree that adding the antidote is not a cause of Victim's survival. Rather, it seems as if Assistant Bodyguard and Buddy are playing a trick on Victim: "we might suppose that Assistant Bodyguard is up for a promotion ... and wants to make it look as though he has foiled an assassination attempt. Buddy is helping him." (Hitchcock, 2007)[p. 520]. The model for this example is $D := \neg A \wedge B$, $B := A$, and the context is such that A holds. It is easy to see that A produces $\neg D$ in the actual story, and that $\neg A$ would likewise produce $\neg D$ in the counterfactual story. This example is thus nothing but a switch, and hence our definition does not consider A a cause of $\neg D$. Looking back at our discussion in Section 7, it is revealing that Weslake – and others with him – judges this example to be similar to Early Preemption, but fails to note the similarity to Switch. We can accommodate for the observation that some people have different intuitions here in the same manner as we did for those examples by pointing out that the backup process is assumed to be completely reliable, which might strike some as unrealistic.

The second example, named "Careful Antidote", is similar in structure to Examples 2 and 3 (Weslake, 2015)[p. 20].[16]

Example 10. *Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Victim's coffee ($\neg A$). Bodyguard puts antidote in the coffee (B), which would have neutralized the poison. Victim drinks the coffee and survives ($\neg D$).*

As with the previous example, adding the antidote intuitively is not a cause of Victim's survival. Once more this spells trouble for many definitions, given the resemblance to symmetric overdetermination, where our intuitions are reversed. We are able to look beyond this resemblance and handle this example as a case of *Late Preemption*, in the same manner as we distinguished between LP and SO, namely by using the timing. Recall that other approaches avoid having an explicit timing by adding additional variables, such as *SH* and BH in LP. Here, there are no obvious candidates for such variables, which explains why they struggle with this example.

Weslake uses the single equation model $D := A \wedge \neg B$. While our approach also gives the correct result for this model, we will explain our reasoning with the following more detailed one: $D := Dr \wedge L$ represents the fact that Victim dies if he drinks a lethal coffee, and $L := A \wedge \neg B$ represents the fact that the coffee is lethal if Assassin poisons it and Bodyguard does not add an antidote. The context is such that Dr, $\neg A$ and B hold.

As $\neg D$ is dependent on $\neg L$, it is clear that $\neg L$ causes $\neg D$. Note also that any causal status of either Assassin or Bodyguard is mediated entirely through L. Since Assassin comes first, from the moment he refrains from putting poison in the coffee, it is not lethal. (Or to be precise, the coffee is no longer potentially lethal, as of course it was not lethal to start with.) Concretely, $\tau(\neg L) = \tau(\neg A) < \tau(B)$. Hence whatever Bodyguard's action might be, it is too late and is not a cause of $\neg D$, in agreement with our intuition.

On the other hand if we change the story so that the order of Assassin and Bodyguard is reversed, then our definition *would* judge B a cause. Indeed, as soon as B happens, i.e.,

[14] One should take into account our discussion of early preemption from Section 7 though: Weslake uses the deterministic model for EP and still judges there to be causation, whereas we claim there is causation only when using the non-deterministic model.

[15] As a notable exception, Hall's account (2007) is able to deal with all of these examples succesfully. (Although he would have to add an extra variable to the model for the Backup example, and he disagrees with Weslake on the trumping causation example). Unfortunately it falls victim to other counterexamples, the most well-known being those from Hitchcock (2009). Again we leave it to the reader to verify that our definition does deliver the right verdict in all of the examples discussed there as well.

[16] An almost identical example is given by Hall (2007), named "back-up threat canceller". He uses it as an example that escapes his earlier dual-concept view of causation as being either dependence or production, and motivated him to develop his later definition. As the analysis will show, our more tolerant notion of production does capture this example. Thus it serves as a good illustration of how our notion of production extends his.

the antidote is added, the coffee has become poison-proof, i.e., no longer potentially lethal. Given that Assassin's omitting the poison happens afterwards, we see that B must be the cause of $\neg L$. Hence it is a cause of $\neg D$ as well.

Lastly, in case we are unable to tell which happened first, $\neg A$ or B, we call both of them causes, just as we did for SO.

To some calling B a cause, even when it happens first, may initially sound counterintuitive (while others may not have any clear intuitions here at all). Given the structural similarities between examples with conflicting, or simply confusing, intuitions, it is too much to expect of any definition that it will align perfectly with intuition in all cases. However an important benefit of our principled account, is that precisely by pointing out the similarities we can show how the same principles are at work in intuitively different examples, and possibly transform people's initial unreflective intuitions into informed judgments.

For example, it could be objected that according to our definition, even if in the end Victim changes his mind, and *does not drink the coffee*, B would nonetheless be a cause of $\neg D$.[17] There is no escaping the fact that initially this sounds counterintuitive. We believe the problem lies with the vagueness surrounding the nature of omissions, and their connection to time.[18] In this example, the omission is the statement "Victim does not die", or perhaps better, "Victim does not die from drinking poisonous coffee". At the start of the example Victim's death had not yet been prevented, and at the end it has. Hence there must be some particular event that happened in between such that Victim's death was prevented precisely at the moment this event occurred. The question is which event? Intuitions seem to be at a loss here, as there is no obvious candidate which presents itself. Certainly refusing to drink a perfectly fine coffee cannot be a cause of Victim's failing to die. We suggest that the way out is by using our principled approach, which generalizes the lessons learned from other examples about which we do have firm intuitions. Therefore the first event which prevented victim's death should be judged its cause.

We come back to the original story to illustrate the vagueness regarding omissions and their timing. The story states the omission that Assassin does not put the poison in Victim's coffee, and that he does not do so because he has a last-minute change of heart. The fact that the statement regarding Bodyguard follows the one regarding Assassin, indicates a temporal order: first, Assassin refrains from putting in poison, then, Bodyguard adds antidote. But intuitively it is not at all clear what it means for Assassin's omission to occur first, precisely because it is not clear what event occurs (and when it does so) such that Assassin's mental state shifts from "intending to put poison in Victim's coffee" to "no longer intending to put poison in Victim's coffee".

This is confirmed if we adapt the example so that we focus only on the timings of events, as with *Late Preemption* and *Symmetric Overdetermination*. Imagine that we start out with a coffee that is already poisonous, and both Assassin and Bodyguard add an effective antidote. In that case our intuitions would simply follow the temporal order by which the antidotes were added: if Assassin adds his first, then Bodyguard adding the antidote is not a cause, and vice versa. We claim that, by analogy, it makes sense to call Assassin's refusal to poison the coffee a cause, as long as he makes up his mind *before* the antidote is put in.

[17] For a very similar example, see "non-existent threats" (Hall, 2007).

[18] Hitchcock (2007) analyses these types of examples using a default/deviant distinction As we mentioned in Section 5, the temporal asymmetry between events and omissions in our notion of a timing can also be interpreted as invoking such a distinction.

12 Conclusion

Our goal in this paper has been to construct a definition of actual causation from the ground up. We have formulated several principles which we take to be fundamental properties of causation, and illustrated each of them by way of a simple example. As a result we derived a definition that is a compromise between the pull of two distinct concepts, namely dependence and production. Given that all three concepts agree on a large number of examples, it is not surprising that the distinction between them is often neglected or misunderstood.

We have applied our definition successfully on a number of paradigmatic examples: symmetric overdetermination, late/early preemption, switching, careful poisoning, and careful antidote. In addition, we have also checked our definition against all examples found in (Hall, 2000, 2004, 2007; Halpern and Pearl, 2005; Halpern, 2016; Hitchcock, 2001, 2007, 2009; Weslake, 2015). Our definition can be applied to all of them along the same lines as we have applied it to the examples mentioned.

We hope our principled approach proves useful as well to those who contest our resulting definition, by clarifying formally how causal judgments depend on accepting or refuting the underlying principles. Further, we believe the interplay between the three concepts here described offers a fruitful perspective for understanding different aspects and interests present in causal stories. In future work we intend to apply this insight by comparing the role of causation in different domains, such as the positive sciences, history, and ethics.

Lastly, our principled definition makes it easier to argue for or against specific causal judgments regarding complex examples. Despite the fact that our definition agrees with intuition in simple paradigmatic cases, we are not forced to seek recourse in intuitions to justify our answers in all cases. Given the diversity of intuitions and their mutual inconsistency, it is essential to have a principled method to settle causal judgments one way or the other.

13 Appendix

Theorem 1. *Given a valid timing τ, E is dependent on C w.r.t. $(M, \vec{u})$ if and only if both of the following conditions hold:*

- *[Condition 1]: C is a producer of E w.r.t. $(M, \vec{u}, \tau)$.*
- *[Condition 2]: $\neg C$ is a producer of $\neg E$ w.r.t. $(M_{do(\neg C)}, \vec{u}, \tau_{do(\neg C)})$.*

Proof. The implication from right to left is trivial, hence we only need to prove the implication from left to right.

Assume E is dependent on C w.r.t. $(M, \vec{u})$, or in other words, $(M, \vec{u}) \models C \wedge E$ and $(M_{do(\neg C)}, \vec{u}) \models \neg E$.

Take τ to be any valid timing w.r.t. $(M, \vec{u})$, $n = \tau(E)$, and $m = \min_{k \in \mathbb{N}}\{L^k_{(M,\vec{u})}$ is sufficient for $E\}$. We first prove that C is a producer of E w.r.t. $(M, \vec{u}, \tau)$.

Take $L^1 \subseteq L^m_{(M,\vec{u})}$ 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 $L^m_{(M,\vec{u})}$ one by one.) By construction, all literals in L^1 are direct actual contributors to E. Moreover, since $m \leq n$, these literals are direct producers of E as well.

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 producer of 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 producer F of D such that $F \notin L_{(M_{do(\neg C),\vec{u}})}$. Since production is transitive, F is a producer of 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 a producer of E. Therefore we conclude that C is a producer of E w.r.t. $(M, \vec{u}, \tau)$.

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

□

References

Beckers S, Vennekens J (2016a) A general framework for defining and extending actual causation using cp-logic. International Journal for Approximate Reasoning 77:105–126

Beckers S, Vennekens J (2016b) The transitivity and asymmetry of actual causation. Ergo

Collins J (2000) Preemptive prevention. Journal of Philosophy 97(4):223–234

Gerstenberg T, Goodman ND, Lagnado DA, Tenenbaum JB (2015) How, whether, why: Causal judgments as counterfactual contrasts. In: Proceedings of the 37th Annual Conference of the Cognitive Science Society, pp 782–787

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

Hall N, Paul LA (2003) Causation and preemption. In: Clark P, Hawley K (eds) Philosophy of Science Today, Oxford University Press

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, Hitchcock C (2015) Graded causation and defaults. The British Journal for the Philosophy of Science 66(2):413–457

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 (2007) Prevention, preemption, and the principle of sufficient reason. The Philosophical review 116(4):495–532

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

Hitchcock C, Knobe J (2009) Cause and norm. Journal of Philosophy 106:587–612

Hume D (1748) An Enquiry concerning Human Understanding

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

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

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

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

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

Vennekens J (2011) Actual causation in cp-logic. Theory and Practice of Logic Programming 11:647–662

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