A General Framework for Defining and Extending Actual Causation using CP-logic

Sander Beckers[1], Joost Vennekens

Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium

Abstract

A central problem in the field of causal modelling is to provide a suitable definition of actual causation, i.e., to define when one specific event caused another. Although current research contains many different definitions, it is pervaded with ambiguities and confusion. Our research has two main goals. First, we wish to provide a clear way to compare competing definitions, and improve upon them so that they can be applied to a more diverse range of instances, including non-deterministic ones. To achieve this we provide a general, abstract definition of actual causation, formulated in the context of the expressive language of CP-logic (Causal Probabilistic logic). We will then show that three recent definitions by Ned Hall (originally formulated for structural models) and a definition of our own (formulated for CP-logic directly) can be viewed and directly compared as instantiations of this abstract definition, which also allows them to deal with a broader range of examples. Second, our framework allows for improving on definitions of actual causation in another way, by incorporating the influence of normality. A recent paper by Halpern and Hitchcock draws on empirical research regarding people's causal judgments, to suggest a graded and context-sensitive notion of actual causation. We rephrase their approach into the probabilistic setting of our abstract definition, allowing us to improve it.

Keywords: Actual Causation, CP-logic, Normality, Norms, Counterfactual dependence

1. Introduction

Suppose we know the causal laws that govern some domain, and that we then observe a story that takes place in this domain; when should we now say that, in this particular story, one event caused another? Ever since Lewis (1973) first analyzed this problem of *actual causation* (a.k.a. token causation) in terms of counterfactual dependence, philosophers and researchers from the AI community alike have been trying to improve on his attempt. The literature now contains many definitions of actual causation, each with its own strengths and weaknesses. The fact that the formulations of these different definitions diverges widely, proves a major obstacle for evaluating and comparing them. A second problem is posed by the standard practice of assuming that questions of causation can be separated from their field of application, because this conflicts with recent research that suggest actual causation is a strongly context-dependent concept. In this paper we address both issues, by presenting a formal framework for expressing definitions and extensions of actual causation.

Following Pearl (2000), structural equations have become a popular formal language for defining actual causation (Woodward, 2003; Hitchcock, 2007, 2009; Halpern and Pearl, 2005a; Weslake, 2015; Halpern, 2015). A notable exception is the work of Ned Hall, who has extensively critiziced the privileged role of structural equations for causal modelling, as well as the definitions that have been expressed with it. He has proposed several definitions himself (Hall and Paul, 2003; Hall, 2004, 2007), the latest of which is a sophisticated attempt to overcome the flaws he observes in those that rely too heavily on structural equations. We have developed a definition of our own (Beckers and Vennekens,

Email addresses: Sander.Beckers@cs.kuleuven.be (Sander Beckers), Joost.Vennekens@cs.kuleuven.be (Joost Vennekens)
[1]Corresponding author

2012; Vennekens, 2011), within the language of CP-logic (Causal Probabilistic logic). CP-logic (Vennekens et al., 2009) is a probabilistic logic programming language, based on Sato's distribution semantics (Sato, 1995).

The relation between these different approaches is currently not well understood. Indeed, they are all expressed using different formalisms (e.g., neuron diagrams, structural equations, CP-logic, or just natural language). Therefore, comparisons between them are limited to verifying on which examples they (dis)agree. Our first goal in this paper is to work towards a remedy for this situation. We will present a general, parametrized, and probabilistic definition of actual causation.

We will develop this framework in the context of CP-logic, because this language offers all the features that are required to define the necessary concepts in a straightforward and natural way. In particular, we make use of the fact that CP-logic has a modular rule-based structure and a semantics that is explicitly temporal and makes a distinction between the default and deviant values of variables. In the context of structural models, which lack these features, our general framework would be significantly more cumbersome to define.

Exploiting the fact that neuron diagrams and structural equations – at least those typically used in the actual causation literature – can be reduced to CP-logic, we will then show that our definition and three definitions by Ned Hall can be seen as particular instantiations of this parametrized definition. Our analysis thus allows for a formal and fundamental comparison between these approaches, which is a first step towards an improved account. Also, it generalises the definitions by Hall from a deterministic to a non-deterministic setting. Still, all of these definitions share one feature: they do not take into account the context in which the question of actual causation is posed.

In their forthcoming article *Graded Causation and Defaults*, Halpern and Hitchcock – HH – quite rightly observe that not only is there a vast amount of disagreement regarding actual causation in the literature, but there is also a growing number of empirical studies which show that people's intuitions are influenced to a large degree by contextual factors which up to now have been ignored when dealing with causation. For example, our judgments on two similarly modelled cases may differ depending on whether it takes place in a moral context or a purely mechanical one, or on what we take to be the default setting, or on whether we take something to be a background condition or not, etc. (Knobe and Fraser, 2008; Moore, 2009; Hitchcock and Knobe, 2009). This has led HH to develop a flexible framework that allows room for incorporating different judgments on actual causation. More specifically, in their view the difference between cases that are modelled using similar structural models depends on which worlds we take to be more normal than others in the different contexts. Therefore their solution is to extend structural models with a normality ranking on worlds, and use it to adapt and order our judgments of actual causation in a manner suited for the particular context.

We sympathize with many of their observations, and we agree that normality considerations do influence our causal judgments. However, we find their representation of normality lacking for three reasons. First, although they emphasize the importance of distinguishing between statistical and normative normality, they use a single ranking for both. Second, they refrain from using probabilities to represent statistical normality, and instead work with a partial preorder over worlds. Third, although they stress the generality of their approach, they develop it solely for the HP-definition of actual causation (Halpern and Pearl, 2005a). The second goal of this paper is to use the general, parametrized definition of actual causation, to improve upon the extension to actual causation offered by HH. Using CP-logic, we are able to represent statistical normality in the usual way, i.e., by means of probabilities. As we will show, such a quantitative representation of statistical normality avoids a number of problems that HH's ordinal representation runs into. To cope with normative normality, we introduce a separate notion of norms. The result will be a more generally applicable and yet simpler approach.

We first introduce the CP-logic language in Section 2. In Section 3, a general definition of actual causation is first presented, and then instantiated into four concrete definitions. Section 4 offers a succinct representation of all these definitions, and an illustration of how they compare to each other. The following section presents the extension to actual causation by HH. We translate their work into the CP-logic language in Section 6. Section 7 contains a first improvement to this translation, followed by some examples and our final extension to actual causation in Section 8.

Part of this work was presented at the Probabilistic Logic Programming workshop (Beckers and Vennekens, 2015b) and at the Commonsense Reasoning Symposium (Beckers and Vennekens, 2015a).

2. CP-logic

We give a short, informal introduction to CP-logic. A detailed description can be found in (Vennekens et al., 2009, 2010). The basic syntactical unit of CP-logic is a CP-law, which takes the general form $Head \leftarrow Body$. The body can in general consist of any first-order logic formula. However, in this paper, we restrict our attention to conjunctions of ground literals. The head contains a disjunction of atoms annotated with probabilities, representing the possible effects of this law. When the probabilities in a head do not add up to one, we implicitly assume an *empty* disjunct, annotated with the remaining probability.

Each CP-law models a specific *causal mechanism*. Informally, if the *Body* of the law is satisfied, then at some point it will be applied, meaning one of the disjuncts in the $Head$ is chosen, each with their respective probabilities. If a disjunct is chosen containing an atom that is not yet **true**, then this law causes it to become **true**; otherwise, the law has no effect. A finite set of such CP-laws forms a CP-theory, and represents the causal structure of the domain at hand. The domain unfolds by laws being applied one after another, where multiple orders are often possible, and each law is applied at most once. We illustrate with an example from (Hall, 2004):

Example 1. *Suzy and Billy each decide to throw a rock at a bottle. When Suzy does so, her rock shatters the bottle with probability* 0.9*. Billy's aim is slightly worse and he only hits with probability* 0.8*.*

This small causal domain can be expressed by the following CP-theory T:

$$Throws(Suzy) \leftarrow . \quad (1) \qquad (Breaks : 0.9) \leftarrow Throws(Suzy). \quad (3)$$

$$Throws(Billy) \leftarrow . \quad (2) \qquad (Breaks : 0.8) \leftarrow Throws(Billy). \quad (4)$$

The first two laws are *vacuous* (i.e., they will be applied in every story) and *deterministic* (i.e., they have only one possible outcome, where we leave implicit the probability 1). The last two laws are *non-deterministic*, causing either the bottle to break or nothing at all.

The given theory summarizes all possible *stories* that can take place in this model. For example, it allows for the story consisting of the following chain of events:

Example 2. *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.*

To formalize this idea, the semantics of CP-logic uses *probability trees* (Shafer, 1996). For this example, one such tree is shown in Figure 1. Here, each node represents a state of the domain, which is characterized by an assignment of truth values to the atomic formulas, in this case $Throws(Suzy)$, $Throws(Billy)$ and $Breaks$. In the initial state of the domain (the root node), all atoms are assigned their *default* value **false**. In this example, the bottle is initially unbroken and the rocks are still in Billy and Suzy's hands. The children of a node x are the result of the application of a law: each edge (x,y) corresponds to a specific disjunct that was chosen from the head of the law that was applied in node x. In this particular case, law (1) is applied first, so the assignment in the child-node is obtained by setting $Throws(Suzy)$ to **true**, its *deviant* value. The third state has two child-nodes, corresponding to law (3) being applied and either breaking the bottle (left child) or not (right child). The leftmost branch is thus the formal counterpart of the above story, where the last edge represents the fact that Billy's throw was also accurate, even though there was no bottle left to break. A branch ends when no more laws can be applied.

A probability tree of a theory T in CP-logic defines an *a priori* probability distribution P_T over all things that might happen in this domain, which can be read off the leaf nodes of the branches by multiplying the probabilities on the edges. For instance, the probability of the bottle breaking is the sum of the probabilities of the leaves in which $Breaks$ is **true** – the white circles in Figure 1 – giving 0.98. We have shown here only one such probability tree, but we can construct others as well by applying the laws in different orders.

An important property however is that all trees defined by the same theory result in the same probability distribution. To ensure that this property holds even when there are bodies containing negative literals, CP-logic makes use of the well-founded semantics. Simply put, this means the condition for a law to be applied in a node is not merely that its body is currently satisfied, but that it will remain so. This implies that a negated atom in a body should not only be currently assigned **false**, but actually has to have become impossible, so that it will remain **false** through to the end-state. For atoms currently assigned **true**, it always holds that they remain **true**, hence here there is no problem.

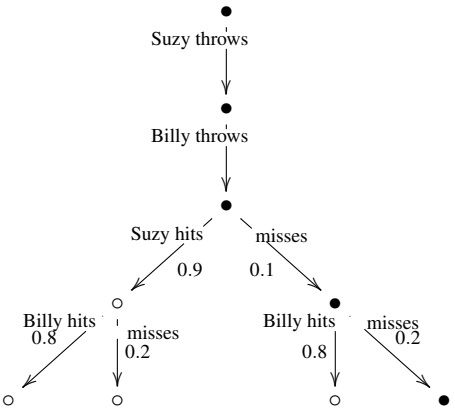

Figure 1: Probability tree for Suzy and Billy.

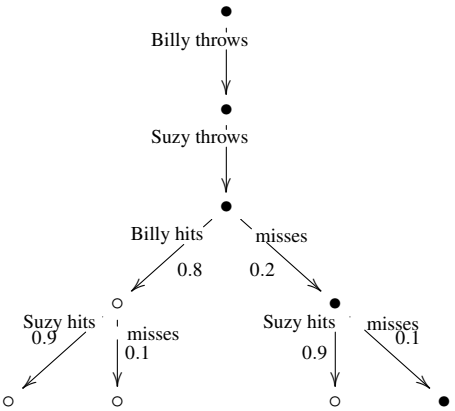

Figure 2: Second probability tree for Suzy and Billy.

The order by which events happen in a causal story, such as Example 2, is represented by the order that laws are applied in a branch. Therefore if the example were slightly different, so that Billy's rock actually hit the bottle first, then the formal counterpart of the story would be the leftmost branch of the probability tree depicted in Figure 2. (Note that if we were to reverse the order of the first two edges, then this would represent the story where Suzy throws before Billy, but Billy throws harder and thus his rock still reaches the bottle first.)

Probability trees do not allow for simultaneous events: at every moment only one event can happen. Although this is a realistic assumption when considering a continuous time-scale, the granularity of the variables of interest in typical examples is usually closer to a coarser, discrete division of time. More concretely, we can easily imagine a variant of the story so that we can't distinguish which rock hits the bottle first. In that case it makes perfect sense to assume that for all intents and purposes both rocks hit the bottle simultaneously, making it a case of symmetric overdetermination. To accommodate this possibility we take such a story to correspond to two branches rather than one, and shall say that C causes E if this holds in either one of the branches considered by themselves.[2] These branches consist of the same nodes, ordered differently (eg., the leftmost branches of the trees in Figures 1 and 2 together represent the story in which Suzy's and Billy's rocks hit the bottle at the same time). For matters of simplicity, throughout this paper we shall assume there is no symmetric overdetermination and a story takes the form of a single branch.

[2]We choose to define causation for these cases in this manner because it agrees with the majority verdict in the literature that both overdetermining events are causes. However one could also define it so that there has to be causation in both branches. In this case, neither event would be considered a cause. If there are more than two simultaneous events, we generalize this reasoning in the straightforward way.

2.1. Counterfactual Probabilities

In the context of structural equations, Pearl (2000) studies counterfactuals and shows how they can be evaluated by means of a syntactic transformation. In their study of actual causation and explanations, Halpern and Pearl (2005b, p. 27) also define counterfactual probabilities (i.e., the probability that some event would have had in a counterfactual situation). Vennekens et al. (2010) present an equivalent method for evaluating counterfactual probabilities in CP-logic, also making use of syntactic transformations.

Assume we have a branch b of a probability tree of some theory T. To make T deterministic *in accordance with the choices made in b*, we transform T into T^b by replacing the heads of the laws that were applied in b with the disjuncts that were chosen from those heads in b. For example, if we take as branch b the story from Example 2, then T^b would be:

$$Throws(Suzy) \leftarrow . \qquad Breaks \leftarrow Throws(Suzy).$$
$$Throws(Billy) \leftarrow . \qquad Breaks \leftarrow Throws(Billy).$$

We will use Pearl's $do()$-operator to indicate an intervention (Pearl, 2000). The intervention on a theory T that ensures variable C remains false, denoted by $do(\neg C)$, removes C from the head of any law in which it occurs, yielding $T|do(\neg C)$. For example, to prevent Suzy from throwing, the resulting theory $T|do(\neg Throws(Suzy))$ is given by:

$$\leftarrow . \qquad (Breaks : 0.9) \leftarrow Throws(Suzy).$$
$$Throws(Billy) \leftarrow . \qquad (Breaks : 0.8) \leftarrow Throws(Billy).$$

Laws with an empty head are ineffective, and can thus simply be omitted. The analogous operation $do(C)$ on a theory T corresponds to adding the deterministic law $C \leftarrow$.

With this in hand, we can now evaluate a Pearl-style counterfactual probability "given that b in fact occurred, the probability that $\neg E$ would have occurred if $\neg C$ had been the case" as $P_{T^b}(\neg E|do(\neg C))$.[3]

3. Defining Actual Causation Using CP-logic

We now formulate a general, parametrized definition of actual causation, which can accommodate several concrete definitions by filling in details that we first leave open. We demonstrate this using definitions by Hall and one by ourselves. For the rest of the paper, we assume that we are given a CP-theory T and an actual story b in which both C and E occurred, and that we are interested in whether or not C caused E.[4] By Con we denote the quadruple (T,b,C,E), and refer to this as a *context*.

3.1. Actual Causation in General

For reasons of simplicity, the majority of approaches (including Hall) only consider actual causation in a deterministic setting. Further, it is taken for granted that the actual values of all variables are given. In such a context, counterfactual dependence of the event E on C is expressed by the conditional: *if* $do(\neg C)$ *then* $\neg E$, where it is assumed that all exogenous variables take on their actual values. In our probabilistic setting, the latter translates into making those laws that were actually applied deterministic, in accordance with the choices made in the story. However, in many examples, the story does not specify the actual value of all exogenous variables. For example, if Suzy is prevented from throwing her rock, then we cannot say what the accuracy would have been had she done so. In CP-logic, this would be represented by the fact that law (3) is not applied. Hence, in a more general setting, it is required only that $do(\neg C)$ makes $\neg E$ possible. In other words, we get a probabilistic definition of counterfactual dependence:

Definition 1 (Dependence)**.** *E is* counterfactually dependent *on C in (T,b) iff $P_{T^b}(\neg E \mid do(\neg C)) > 0$.*

[3] Fenton-Glynn (2015) uses such counterfactual probabilities to extend the definition of causation from Halpern and Pearl (2005a) to probabilistic structural equations. His focus lies with incorporating the idea that causes are "probability-raising" with regards to their effect, into a counterfactual account. In this manner, two traditional approaches to actual causation are combined. This project stands somewhat orthogonal to our current investigation. In future work it would be interesting to see how his results can be integrated into our framework.

[4] We hereby limit ourselves to causation between literals C and E, but our account can be generalised to include complex causes and effects as well.

As counterfactual dependency lies at the heart of causation for all of the approaches we are considering, Dependence represents the most straightforward definition of actual causation.[5] It is however too crude and allows for many counterexamples, preemption being the most famous.

More refined definitions agree with the general structure of the former, but modify the theory T in more subtle ways than T^b does. We identify two different kinds of laws in T, that should each be treated in a specific way.

The first are the laws that are *intrinsic* with respect to the given context. These are laws whose outcome is fixed, in the sense that in any counterfactual story we might consider, they will always produce the same outcome as they did in the actual story. Thus, intrinsic laws should be made deterministic in accordance with b.

The second are laws that are *irrelevant* in the given context. These are laws that played no part in the causal process that caused E, and that we should therefore not take into account when trying to find out if C was a cause of E or not. Thus, irrelevant laws should simply be ignored.

Together, the methods of determining which laws are intrinsic and irrelevant, respectively, will be the parameters of our general definition. Suppose we are given two functions Int and Irr, which both map each context (T,b,C,E) to a subset of the theory T. With these, we define actual causation as follows:

Definition 2 (Actual causation given Int and Irr)**.** *Given the context Con, we define that C is an actual cause of E if and only if E is counterfactually dependent on C when replacing T^b with the theory T^* that we construct as:*
$T^* = [T \setminus (Irr(Con) \cup Int(Con))] \cup Int(Con)^b$.

For instance, the naive approach that identifies actual causation with counterfactual dependence corresponds to taking Irr as the constant function $\{\}$ and $Int(Con)$ as $\{r \in T \mid r \text{ was applied in } b\}$. From now on, we use the following, more legible notation for a particular instantiation of this definition:

Dependence-Irr. *No law r is irrelevant.*

Dependence-Intr. *A law r is intrinsic iff r was applied in b.*

The following straightforward theorem expresses that this formulation of dependence is equivalent to the original in Definition 1.

Theorem 1. *Given the context Con, C is an actual cause of E given* Dependence-Irr *and* Dependence-Intr *iff E is counterfactually dependent on C.*

If desired, we can order different causes by their respective counterfactual probabilities $P_{T^*}(\neg E \mid do(\neg C))$, as this indicates how important the cause was for E. Note however that Definition 1 reduces to a standard deterministic definition of counterfactual dependence if all CP-laws are deterministic. In that case, our general Definition 2 becomes deterministic as well.

3.2. Beckers and Vennekens 2012 Definition

A recent proposal by the current authors for a definition of actual causation was originally formulated in (Vennekens, 2011), and later slightly modified in (Beckers and Vennekens, 2012). Here, we summarize the basic ideas of the latter, and refer to it as *BV12*. We reformulate this definition in order to fit into our framework of Definition 2. It is easily verified that both versions are equivalent.

Because we want to follow the actual story as closely as possible, the condition for intrinsicness is exactly like before: we force all laws that were applied in b to have the same effect as they had in b.

To decide which laws were relevant for causing E in our story, we start from a simple temporal criterion: every law that was applied after the effect E took place is irrelevant, and every law that was applied before isn't. For example, to figure out why the bottle broke in our previous example, law (4) is considered irrelevant, because the bottle was already broken by the time Billy's rock arrived. For laws that were not applied in b, we distinguish laws that could still be applied when E occurred, from those that could not. The first are considered irrelevant, whereas the second aren't. This ensures that any story b' that is identical to b up to and including the occurrence of E provides the same judgements about the causes of E, since any law that is not applied in b but is applied in b', must obviously occur after E.

[5] This definition is similar in spirit to that of a *partial explanation* given in (Halpern and Pearl, 2005b). There the probability measures the *goodness* of the explanation, here it measures the *importance* of the cause.

BV12-Irrelevant. *A law r is irrelevant iff r was not applied before E in b, although it could have. (I.e., it was not impossible at the time when E occurred.)*

BV12-Intrinsic. *A law r is intrinsic iff r was applied in b.*

The following theorem expresses that the current formulation is equivalent to the original definition.

Theorem 2. *Given the context Con, C is an actual cause of E given* BV12-Irr *and* BV12-Intr *iff C is an actual cause of E as defined in (Beckers and Vennekens, 2012).*

3.3. Hall 2007

One of the currently most refined concepts of actual causation is that of Hall (2007). Although Hall uses structural equations as a practical tool, he is of the opinion that intuitions about actual causation are best illustrated using neuron diagrams. A key advantage of these diagrams, which they share with CP-logic, is that they distinguish between the default and deviant state of a variable. Proponents of structural equations, on the other hand, countered Hall's approach by criticizing neuron diagrams' limited expressivity (Hitchcock, 2009, p. 398). Indeed, a neuron diagram, and thus Hall's approach as well, is very limited in the kind of examples it can express. In particular, neuron diagrams can only express deterministic causal relations, and they lack the ability to directly express *causation by omission*, i.e., that the absence of C by itself causes E, as in the law $E \leftarrow \neg C$. Hall's solution is to argue against causation by omission altogether. By contrast, we will offer an improvement of Hall's account that generalizes to a probabilistic context, and can also handle direct causation by omission.

3.3.1. Neuron Diagrams, Structural Equations, and CP-logic

In a standard neuron diagram, a neuron can be in one of two states, the default "off" state and the deviant "on" state in which the neuron "fires". Different kinds of links define how the state of one node affects the other. For instance, in Diagram 1, E fires iff at least one of B or D fires, D fires iff C fires, and B fires iff A fires and C doesn't fire. Nodes that are "on" are represented by full circles and nodes that are "off" are shown as empty circles.

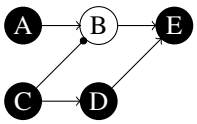

Figure 3: Diagram 1

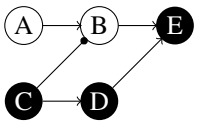

Figure 4: Diagrams 2

Diagrams 1 and 2 represent the same causal structure, but different stories: in both cases there are two causal chains leading to E, one starting with C and another starting with A. But in Diagram 1 the chain through B is preempted by C, whereas in Diagram 2 there is nothing for C to preempt, as A doesn't even fire. Therefore the first is an example of what is generally known as Early Preemption, whereas the second is not.

Although Hall presents his arguments using neuron diagrams, his definition is formulated in terms of structural equations that correspond to such diagrams in the following way: for each endogenous variable there is one equation, which contains a propositional formula on the right-hand side concisely expressing the dependencies of the diagram. The left-most neurons, which have no incoming edges, are determined directly by the exogenous variables $\mathbf{U}$ that represent the background conditions. For example, the structural equations for Diagrams 1 and 2 are:

$$A := U_1. \qquad B := A \wedge \neg C. \qquad E := B \vee D.$$
$$C := U_2. \qquad D := C.$$

A formal description of structural equations can be found in (Pearl, 2000). 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. Therefore, in this article, we restrict attention to those kinds of models. (We point out that CP-logic can also be generalised to include discrete variables with a finite domain as well.) We can translate such a structural model into an equivalent CP-logic

theory by simply replacing the ":=" symbol by "←", and replacing the exogenous variables by non-determinism in the laws (Vennekens et al., 2009).

CP-logic allows representations of causal relations that are more refined than those of structural models in two ways. First, cyclic causal relations can be represented in a more correct way than considered in, e.g., Pearl (2000) — for a discussion of this, see Vennekens et al. (2010)[Section 5]. Second, CP-logic also takes a modular approach to representing causal mechanisms: rather than having a single CP-law that combines all direct causes of a variable, CP-logic splits up independent mechanisms into separate CP-laws (Vennekens et al., 2010). Analogously, we translate a single structural equation into a set of independent causal mechanisms. The same applies to neuron diagrams, which is why CP-logic splits up the influence of B and D on E into two separate laws, similar to the influence of Billy and Suzy on the bottle shattering in Example 1. Concretely, the translation of the causal model from Diagrams 1 and 2 into CP-logic is given by the following CP-theory – where p and q represent some probabilities:

$$(A : p) \leftarrow . \qquad B \leftarrow A \wedge \neg C. \qquad E \leftarrow B.$$
$$(C : q) \leftarrow . \qquad D \leftarrow C. \qquad E \leftarrow D.$$

The state of the neuron diagram (i.e., which variables fire and which do not) corresponds to an assignment $\mathbf{v}$ of values to the boolean variables $\mathbf{V}$ of the corresponding structural equations model M, which in turn corresponds to an interpretation for a vocabulary of the corresponding CP-logic theory. With such a state, we therefore associate the set of all branches of probability trees of the CP-logic theory whose leaf nodes contain this interpretation. Since we stick to cases that do not involve symmetric overdetermination, we can choose a single branch b out of this set to represent a story in a concise fashion.

3.3.2. Hall's Definition

The idea behind Hall's definition is to check for counterfactual dependence in situations which are reductions of the actual situation, where a reduction is understood as "a variant of this situation in which *strictly fewer* events occur". In other words, because the counterfactual dependence of E on C can be masked by the occurrence of events which are extrinsic to the actual causal process, we look at all possible scenario's in which there are less of these extrinsic events. Hall puts it like this (2007)[p. 129]:

> Suppose we have a causal model for some situation. The model consists of some equations, plus a specification of the actual values of the variables. Those values tell us how the situation *actually* unfolds. But the same system of equations can also represent *nomologically possible variants*: just change the values of one or more exogenous variables, and update the rest in accordance with the equations. A good model will thus be able to represent a range of variations on the actual situation. Some of these variations will be – or more accurately, will be modeled as – *reductions* of the actual situation, in that every variable will either have its actual value or its default value. Suppose the model has variables for events C and E. Consider the conditional
>
> **if C = 0; then E = 0**
>
> This conditional may be true; if so, C is a cause of E. Suppose instead that it is false. Then C is a cause of E iff there is a reduction of the actual situation according to which C and E still occur, and in which this conditional is true.

Rather than speaking of fewer events occuring, in this definition Hall characterizes a reduction in terms of whether or not variables retain their actual value. This is because in the context of neuron diagrams, an event is the firing of a neuron, which is represented by a variable taking on its deviant value, i.e., the variable *becoming true*. In the dynamic context of CP-logic, the formal object that corresponds most naturally to Hall's informal concept of an event is the transition in a probability tree (i.e., the application of a causal law) that makes such a variable true. Therefore we take a reduction to mean that no law is applied such that it makes a variable true that did not become true in the actual setting.

To make this more precise, we introduce some new formal terminology. Let d be a branch of a probability tree of the theory T. $Laws_d$ denotes the set of all laws that were applied in d. The resulting effect of the application of a law

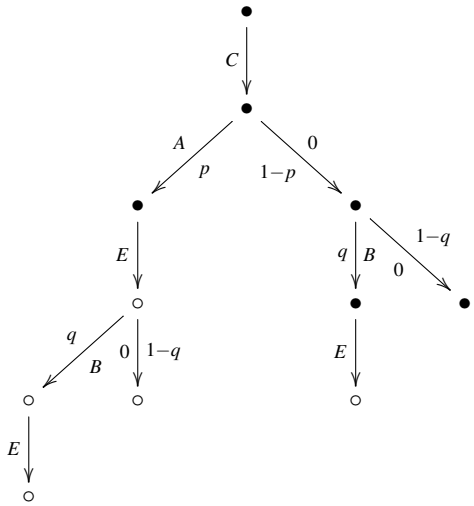

Figure 5: Probability tree with a non-simple story.

$r \in Laws_d$ – i.e., the disjunct of the head which was chosen – will be denoted by r_d, or by 0 if an empty disjunct was chosen. The set of true variables in the leaf of d will be denoted by $Leaf_d$.

A branch d is a *reduction* of b iff $\forall r \in Laws_d : r_d = 0 \vee \exists s \in Laws_b : r_d = s_b$. Or, equivalently, $Leaf_d \subseteq Leaf_b$.

A reduction of b in which both C and E occur – i.e., hold in its leaf – will be called a (C,E)-reduction. The set of all of these will be denoted by $Red_b^{(C,E)}$. These are precisely the branches which are relevant for Hall's definition.

Definition 3. *We define that C is an* actual cause *of E iff* $(\exists d \in Red_b^{(C,E)} : P_{T^d}(\neg E | do(\neg C)) > 0)$.

Theorem 3 shows the correctness of our translation. Proofs of all theorems can be found in the Appendices.

Theorem 3. *Given a neuron diagram with its corresponding equations M, and an assignment to its variables* **V**. *Consider the CP-logic theory T and story b that we get when applying the translation from Section 3.3.1. Then C is an actual cause of E in the diagram according to Hall's definition iff C is an actual cause of E in b and T according to Definition 3.*

At first sight, Definition 3 does not fit into the general framework we introduced earlier, because of the quantifier over different branches. However, we will now show that for a significant group of cases it actually suffices to consider just a single T^*, which can be described in terms of irrelevant and intrinsic laws.

Rather than looking at all of the reductions separately, we single out a minimal structure which contains the essence of our story. This structure will be based on the set of all laws that are *necessary* for a reduction, in the sense that they happen in each reduction and that, moreover, they always happen in the same way, i.e., with the same outcome.

Definition 4. *A law r is* necessary *iff*

- $\forall d \in Red_b^{(C,E)} : r \in Laws_d$ *and*
- $\forall d, e \in Red_b^{(C,E)} : r_d = r_e$.

We define $Nec(b)$ *as the set of all necessary laws.*

In general it might be that there are two (or more) laws which are unnecessary by themselves, but at least one of them has to be applied as it was in b. Consider for example the following CP-theory.

$$C \leftarrow . \qquad (A:p) \leftarrow C. \qquad E \leftarrow A.$$
$$(B:q) \leftarrow C. \qquad E \leftarrow B.$$

In the story where C causes both A and B, each of those being sufficient for E, neither the second nor the third law is necessary for E. Yet it is clear that at least one of them has to be applied to get E. In cases where this complication does not arise, we shall say that the story is simple.

First, some helpful terminology: an r-variant of a branch b is any branch b' which coincides with b up to the application of r, but which then selects a different atom $r_{b'} \neq r_b$ from the head of r. Note that if r is deterministic, there are no r-variants.

Definition 5. *A story b is* simple *iff the following holds:*

- $\forall r \in Laws_b$: *the head of r contains at most two disjuncts;*
- $\forall d \in Red_b^{(C,E)}$, *for all non-deterministic $r \in Laws_d \setminus Nec(b)$: $\exists e \in Red_b^{(C,E)}$ so that e is an r-variant of d.*

We illustrate this concept by looking at a probability tree for the previous example in Figure 5, to show that the previous story is not simple. The left-most branch is a formal counterpart of this story. Except for the right-most branch in which E is **false**, all branches are (C,E)-reductions of b. To see that the second law r is not necessary, observe that for any branch d in the left-side of the tree, $r_d = A$, whereas for any branch e in the right-side, $r_e = 0$. Similarly, the third law is not necessary either.

Now consider the second branch from the right, d, and the third law, r', that results in B in the fourth node: there does not exist a (C,E)-reduction e of b that is identical to d up to the third node (i.e., up to the application of r'), but different regarding the fourth node (i.e., $B = r'_d \neq r'_e$). Hence, b is not simple.

We are now in a position to formulate a theorem that will allow us to adjust Hall's definition into our framework.

Theorem 4. *If $(\exists d \in Red_b^{(C,E)} : P_{T^d}(\neg E|do(\neg C)) > 0)$ then $P_{T^{Nec(b)}}(\neg E|do(\neg C)) > 0$. If b is simple, then the reverse implication holds as well.*

It is possible to add an additional criterion to turn this theorem into an equivalence that also holds for non-simple stories. We choose not to do this, because all of the examples Hall discusses are simple, as are all of the classical examples discussed in the literarure, such as Early and Late Preemption, Symmetric Overdetermination, Switches, etc.

As a result of this theorem, rather than having to look at all (C,E)-reductions and calculate their associated probabilities, we need only find all the necessary laws and calculate a single probability. If the story b is simple, then this probability represents an extension of Hall's definition, since they are equivalent if one ignores the value of the probability but for it being 0 or not. To obtain a workable definition of actual causation, we present a more constructive description of necessary laws.

Theorem 5. *If b is simple, then a non-deterministic law r is necessary iff none of the r-variants of b is a (C,E)-reduction.*

With this result, we can finally formulate our version of Hall's definition, which we will refer to as Hall07.

Hall07-Irrelevant. *No law r is irrelevant.*

Hall07-Intrinsic. *A law r is intrinsic iff r was applied in b, and none of the r-variants d of b is such that $\{C,E\} \subseteq Leaf_d \subseteq Leaf_b$.*

3.4. Hall 2004 Definitions

Hall (2004) claims that it is impossible to account for the wide variety of examples in which we intuitively judge there to be actual causation by using a single, all-encompassing definition. Therefore he defines two different concepts which both deserve to be called forms of causation but are nonetheless not co-extensive.

3.4.1. Dependence

The first of these is simply Dependence, as stated in Definition 1. As mentioned earlier, Hall only considers deterministic causal relations, and thus the probabilistic counterfactual will either be 1 or 0.

3.4.2. Production

The second concept tries to express the idea that to cause something is to bring it about, or to *produce* it. The original, rather technical, definition can be found in the appendices, but the following informal version suffices for our purposes: C is a producer of E iff there is a directed path of firing neurons in the diagram from C to E. In our framework, this translates to the following.

Production-Irr. *A law r is irrelevant iff r was not applied before E in b, or if its effect was already* **true** *when it was applied.*

Production-Intr. *A law r is intrinsic iff r was applied in b.*

Theorem 6. *Given a neuron diagram with its corresponding equations M, and an assignment to its variables* $\mathbf{V}$*. Consider the CP-logic theory T, and a story b, that we get when applying the translation from Section 3.3.1. C is a producer of E in the diagram according to Hall iff C is an actual cause of E given* Production-Irr *and* Production-Intr.

Besides providing a probabilistic extension, the CP-logic version of production also offers a way to make sense of causation by omission. That is, just as with all of the definitions in our framework in fact, we can extend it to allow negative literals such as $\neg C$ to be causes as well.

4. Comparison

Table 1 presents a schematic overview of the four definitions we have discussed. The columns and rows give the criteria for a law r of T to be considered intrinsic, respectively irrelevant, in relation to a story b, and an event E. By $r \leq_b E$, we denote that r was applied in b before E occurred.

Table 1: Spectrum of definitions

Irrelevant	Intrinsic	
	$r \in Laws_b$	$r \in Nec(b)$
$\emptyset$	Dependence	Hall07
$\exists d : (d = b \text{ up to } E) \wedge r \geq_d E$	BV12	
$r \not<_b E \vee r_b <_b r$	Production	

Looking at this table, we can informally characterise the different definitions by describing which events are allowed to happen in the counterfactual worlds they take into consideration to judge causation:

- **Production**: Only those events – i.e., applications of laws making a variable **true** – which happened before E, and not differently – i.e., with the same outcome as in the actual story.
- **BV12**: Those events which happened before E, and not differently, and also those events which were prevented from happening by these.
- **Hall07**: All events, as long as those events that were necessary to E do not happen differently.
- **Dependence**: All events, as long as those events that did actually happen do not happen differently.

In order to illustrate the working of the definitions and their differences, we present an example:

Example 3. *Assassin decides to poison the meal of a victim, who subsequently Dies right before dessert. However, Murderer decided to murder the victim as well, so he poisoned the dessert. If Assassin had failed to do his job, then Backup probably would have done so all the same.*

The causal laws that form the background to this story are give by the following theory:

$$(Assassin : p) \leftarrow . \qquad Dies \leftarrow Assassin.$$
$$(Murderer : q) \leftarrow . \qquad Dies \leftarrow Backup.$$
$$(Backup : r) \leftarrow \neg Assassin. \qquad Dies \leftarrow Murderer.$$

In this story, did *Assassin* cause *Dies*? We leave it to the reader to verify that in this case the left intrinsicness condition from the table applies to the first two non-deterministic laws, whereas the right one only applies to the first. The second irrelevance condition only applies to the last law, whereas the third one applies to the last two laws and to the third. This results in the following probabilities representing the causal status of *Assassin*:

Production	BV12	Hall07	Dependence
1	$1-r$	$(1-r)*(1-q)$	0

Hence *Assassin* is a full cause according to Production, not a cause at all according to Dependence, and somewhere in between these two extremes according to the other two definitions.

Intuitively, most people would judge *Assassin* to be fully responsible for causing victim's death. Hence this particular example seems to speak in favour of Production. However, note that this example is clearly set in a normative context, since murdering people is – in almost all cases – judged to be wrong. One can easily come up with morally neutral examples using these CP-laws and the same story such that our intuitions would be different, for instance the following story:

Example 4. *Billy has set the alarm for six o'clock, at which time it goes off, so that he and Suzy make it in time to school. However, Suzy had put her alarm for five past six, which would have also left ample amount of time. If Billy had failed to put his alarm, then Mother probably have done so all the same.*

In this story, it sounds quite reasonable to say that *Billy* is not a full cause of *Billy* and *Suzy* making it to school on time. We deliberately first chose an example that contains normative elements, because it is a general feature of all existing definitions of causation that they fail to do justice to such context-dependency. We now turn to the second goal of this paper: to incorporate the influence of norms into judgments of actual causation.

5. The HH Extension to Actual Causation

In this section we succinctly present the graded, context-dependent approach to actual causation from (Halpern and Hitchcock, 2015). We will use the following story from (Knobe and Fraser, 2008) as our running example, as it illustrates the influence normative considerations can have on our causal attributions:

Example 5. *The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens. On Monday morning, one of the administrative assistants encounters professor Smith walking past the receptionists desk. Both take pens. Later, that day, the receptionist needs to take an important message...but she has a problem. There are no pens left on her desk.*

We formally represent the relevant events on the Monday morning from our running example using CP-logic. The domain consists of the variables $Prof$ and $Assistant$, which stand for the professor respectively the assistant taking a pen, and $NoPens$, which is true when there are no pens left. The causal structure can be represented by the following CP-theory T:

$$(Prof : 0.7) \leftarrow . \tag{5}$$
$$(Assistant : 0.8) \leftarrow . \tag{6}$$
$$NoPens \leftarrow Prof \wedge Assistant. \tag{7}$$

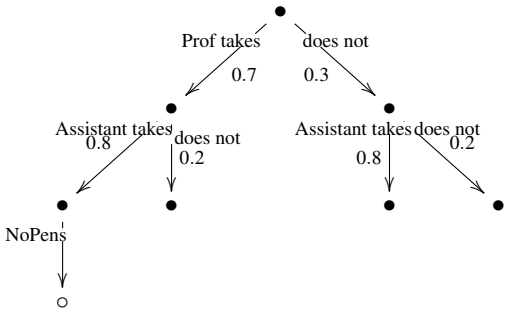

Figure 6: Probability tree for the Pen-vignette.

The given theory summarizes all possible stories that can take place in this model. One of those is what in fact did happen that Monday morning: both the professor and the assistant take a pen, leaving the receptionist faced with no pens. The other stories consist in only the professor taking a pen, only the assistant doing so, or neither, as can be seen in the probability tree in Figure 6. The leftmost branch is the formal counterpart of the above story.

As with much work on actual causation, HH frame their ideas using structural equation modelling. As mentioned earlier, such a model consists of a set of equations, one for each endogenous variable, that express the functional dependencies of the endogenous variables on others. As before, these models contain two types of (boolean) endogenous variables: the ones that deterministically depend on other endogenous variables, and those that depend directly on exogenous variables.

HH take a *world* to be an assignment to all endogenous variables. Given a structural model, each assignment $\mathbf{u}$ to the exogenous variables determines a unique world, denoted by $s_{\mathbf{u}}$.

An extended structural model $(M, \succeq)$ consists of a structural model M together with a normality ranking $\succeq$ over worlds. This ranking is a partial pre-order informed by our – possibly subjective – judgments about what we take to be normal in this context. It is derived by considering the typicality of the values of the variables in each world. A world s is more normal than s' if there is at least one variable that takes a more typical value in s than it does in s', and no variable takes a less typical value. For the variables that depend only on other endogenous variables, things are straightforward: it is typical that these variables take the value dictated by their deterministic equation. For the other variables, i.e., those that depend only on the exogenous variables, a typicality ranking over their possible values has to be provided by the modeller.

Typicality and normality are meant to encompass both statistical and normative judgments. Although HH make no syntactic distinction between the two kinds of normality, in the examples discussed they do differentiate between them informally. By contrast, we will make a formal distinction between the two, because in this manner we can incorporate information regarding both.

Say we have a story, i.e., an assignment to all variables, such that C and E happen in it. In order to establish whether C is a cause of E, any definition in the counterfactual tradition restricts itself to some particular set of counterfactual worlds in which $\neg C$ holds and checks whether also $\neg E$ holds in these worlds. If this set contains a world which serves to justify that C is indeed a cause of E (i.e., one in which E is false), then HH call such a world a *witness* of this. HH adapt a given definition of actual causation using the normality ranking to disallow worlds that are less normal than the actual world, in order to reflect the influence of normality on possible causes.

Definition 6. *[HH-extension of actual causation.] Given are an extended structural model $(M, \succeq)$ and exogenous assignment $\mathbf{u}$, such that both C and E hold in $s_{\mathbf{u}}$. C is an HH-actual cause of E in $(M, \succeq, \mathbf{u})$ iff C is an actual cause of E in $(M, \mathbf{u})$ when we consider only witnesses w such that $w \succeq s_{\mathbf{u}}$.*

Since we have a ranking on the normality of worlds, this definition straightforwardly leads to an ordering between different causes indicating the strength of the causal relationship by looking at the highest ranked witness for a cause, which is called its *best witness*.

In case of our example, we get that

$$\{\neg Prof, Assistant, \neg NoPens\} \succ \{Prof, Assistant, NoPens\} \succ \{Prof, \neg Assistant, \neg NoPens\}$$

Since the actual world is in the middle, the first of these can serve as a witness for $Prof$ being a cause, but the last may not be used to judge $Assistant$ to be a cause.

6. The HH Extension in CP-logic

We proceed with translating the HH-extension of actual causation into CP-logic. To get there, we will translate one by one all of the required concepts.

HH use the HP-definition (Halpern and Pearl, 2005a) as their working definition to illustrate their extension to actual causation. However, they stress the generality of their approach, and mention that one could for example apply it to Hall's definition (2007), which we reformulated as an instantation of our general definition in Section 3.3. To keep things simple, we will use Dependence as our definition of actual causation throughout the examples, and thus take T^* to simply be T^b. Note however that our approach can be applied to any instantiation of our general framework.

A structural model M in the HH-setting corresponds to a CP-theory T: a direct dependency on the exogenous variables results in a non-deterministic, vacuous law (such as (5) and (6)), while a dependency on only endogenous variables results in a deterministic law (such as (7)).

A world s described by M then corresponds to a branch b – or to be more precise, the leaf of a branch – of a probability tree of T. In Definition 6, we restricted attention to those worlds that are at least as normal as the actual world. The CP-logic-equivalent of this will be the *normal refinement* of T according to b, which is a theory that describes those stories which are at least as normal as b.

First we introduce two operations on CP-laws, corresponding to the two different interpretations of normality. Assume at some point a CP-law r was applied in b, choosing the disjunct r_b that occurs in its head.

On a probabilistic reading, an alternative application of r is at least as normal as the actual one if a disjunct is chosen which is at least as likely as r_b. Therefore the *probabilistically normalized* refinement of r according to b – denoted by $r^{PN(b)}$ – consists in r without all disjuncts that have a strictly smaller probability than r_b. Remain the laws that were not applied in b. In line with the understanding of normality from HH, we choose to handle these such that they cannot have effects which result in a world less normal than the actual world. Thus we remove those disjuncts that are false in the leaf of b and have a probability lower than 0.5. Further, say the total probability of the removed disjuncts in some law is p, then we renormalize the remaining probabilities by dividing by $1-p$. (Unless $p=1$, then we simply remove the law.)

Definition 7. *Given a theory T, and a story b, we define the* probabilistically normalized refinement *of T according to b as $T^{PN(b)} := \{r^{PN(b)} | r \in T\}$.*

In case of our example, $T^{PN(b)} = T^b$, shown earlier.

A second reading of normality considers not what did or could happen, but what ought to happen. To allow such considerations, we extend CP-logic with *norms*. Everything that can possibly happen is described by the CP-laws of a theory, hence we choose to introduce *prescriptive* norms in addition to *descriptive* CP-laws. These take the form of alternative probabilities for the disjuncts in the head of a law, which represent how the law should behave. (A more general approach could be imagined, but for the present purpose this extension will suffice.) These probabilities will be enclosed in curly braces, and have no influence on the actual behaviour of a theory. However, if we wish to look at how the world should behave, we can enforce the norms by replacing the original probabilities of a theory T with the normative ones. For example, extending law (5) with the norm that the professor shouldn't take pens, even though he often does, gives $(Prof : 0.7\{0\}) \leftarrow$. The *normatively normalized* refinement – denoted by $r^{NN(b)}$ – is then given by the CP-law $(Prof : 0) \leftarrow$. To properly capture the HH definition, our normalized theory should allow all worlds that are at least as normal as b, including of course b itself. For this reason, we here restrict attention to norms with a probability p so that $0 < p < 1$, because a norm $p = 0$ or $p = 1$ could make the actual world b impossible. This restriction is lifted in our own proposal in Section 8.1. In the meanwhile we shall use $(Prof : 0.7\{p\}) \leftarrow$, where p is some small probability, e.g., 0.01.

Definition 8. *Given an extended theory T, i.e., a theory also containing norms, we define the* normatively normalized refinement *of T as $T^{NN} := \{r^{NN} | r \in T\}$.*

We can combine both senses of normality, as follows:

Definition 9. *Given an extended theory T, and a story b, we define the* normal refinement *of T according to b as* $T^{Normal(b)} := (T^{NN})^{PN(b)}$.

The normal refinement according to b is constructed out of T by eliminating all the disjuncts with values of variables that are less normal (either probabilistically or normatively) than the values from b, and thus it allows precisely those stories which are at least as normal as b.[6] In case of our example, $T^{Normal(b)}$ is given by:

$$(Prof : p) \leftarrow . \tag{8}$$

$$Assistant \leftarrow . \tag{9}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{10}$$

To be able to prove that this is indeed the same as HH, we still need to explain how we get from a given extended structural model to an extended CP-theory. The normality ranking is derived by considering what is typical for those variables depending directly on the exogenous variables, i.e., those that we represent by $X : p \leftarrow$. HH use statements that take the form: "it is typical for the variable X to be **true**", or "it is typical for it to be **false**". In CP-logic this becomes: $p > 0.5$, and $p < 0.5$ respectively. A statement of the form: "it is more typical for X to be **true** than for Y to be **true**" translates in an ordering on the respective probabilities. If the typicality statement is of the normative kind, then it is best represented by norms in CP-logic. Thus if there is a norm regarding X then the law will take the extended form $X : p\{q\} \leftarrow$.

In Definition 6, we have that a world is an acceptable witness only if it belongs to the set of worlds allowed by the definition of actual causation that is being used, and it is at least as normal as the actual world. Similarly, we need to limit the stories allowed by Definition 2 – which are described by $T^*|do(\neg C)$ – to those stories which are at least as normal as b. We would like to do this in exactly the same manner as we did for T, i.e., by looking at $(T^*|do(\neg C))^{Normal(b)}$. However, by applying the intervention $do(\neg C)$, which removes C from the head of any law in T in which it appears, we lose all information on the (ab)normality of $\neg C$. As this information should be taken into account as well, we work with the normal refinement of these laws instead. Concretely, denote by $R(C)$ those laws from T which have C in their head. We denote by T^{**} the theory identical to $T^*|do(\neg C)$ except that it contains the normal refinements of all laws in $R(C)$. This complication disappears when we consider our improvement to the HH approach further on.

Definition 10. *Given an extended theory T, a story b such that C and E hold in its leaf, and the theory T^* as described in Definition 2. We define the* normal refinement *of T^* according to b and C as* $T^{Normal(b)*} := (T^{**})^{Normal(b)}$.

This leads us to the following formulation of the HH-approach in CP-logic.

Definition 11. *[HH-CP-logic-extension of actual causation] Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an* HH-CP-logic-actual cause *of E in (T,b) iff* $P_{T^{Normal(b)*}}(\neg E \wedge \neg C) > 0$.

For $C = Assistant$, the probabilistic normalisation of T^{**} replaces $(Assistant : 0.8) \leftarrow$ with the deterministic law $Assistant \leftarrow$, leading to $P(\neg NoPens \wedge \neg Assistant) = 0$. On the other hand, for $C = Prof$, the normative normalisation replaces $(Prof : 0.7\{p\}) \leftarrow$ with $(Prof : p) \leftarrow$, leading to $P(\neg NoPens \wedge \neg Prof) = 1 - p$. Thus $Prof$ is judged to be a strong cause of $NoPens$, whereas $Assistant$ isn't a cause at all, in line with the empirical results from (Knobe and Fraser, 2008). Note that it is only by using the normative probabilities rather than the statistical ones that we get the correct response for $Prof$.

In Section 3.1 we gave a general definition of actual causation in terms of CP-logic. In order to finish the translation from extended structural models to extended CP-logic, we assume that the definition of actual causation being used is such that just as the three definitions from Hall discussed in Section 3.3 and 3.4, it can be translated from structural models into the framework of our general definition. We now show that Definition 11 is indeed the correct translation of the HH approach from structural models to CP-logic.

Theorem 7. *C is an HH-actual cause of E in an extended model and exogenous assignment $(M, \succeq, \mathbf{u})$ iff C is an HH-CP-logic-actual cause of E in (T,b), where (T,b) is derived from $(M, \succeq, \mathbf{u})$ in the sense described above.*

[6] A formal proof of this can be found in the Appendices, as Lemma 2.

7. The Importance of Counterfactuals

We mentioned earlier that one criterion for a story to be normal was that it respects the laws/equations. On the other hand definitions of actual causation look at counterfactual stories resulting from an intervention, namely $do(\neg C)$, which violates the laws. Following HH, Definition 11 tries to circumvent the use of this intervention by only allowing witnesses in which $\neg C$ happens to hold. However, it may be the case that this condition actually eliminates all potential witnesses. When this happens, counterintuitive results may follow. We illustrate what goes wrong by using the following theory:

$$(A : 0.1) \leftarrow . \qquad E \leftarrow C.$$
$$C \leftarrow A. \qquad E \leftarrow \neg A.$$

Consider the story where first A occurs, followed by C and E. Intuitively, C is a strong cause of E, because it is an atypical phenomenon ($P(C) = 0.1$) without which E would not have occured. The law with A in its head is intrinsic, and thus $T^{Normal(b)*}$ is:

$$A \leftarrow . \qquad E \leftarrow C.$$
$$C \leftarrow A. \qquad E \leftarrow \neg A.$$

Applying the definition, we get that $P_{T^{Normal(b)*}}(\neg E \wedge \neg C) = 0$, giving the absurd result that C is not a cause of E at all. The problem lies in the fact that in its current form we only allow stories containing $\neg C$ in the usual, lawful way, rather than stories which contain $\neg C$ as a result of the intervention $do(\neg C)$. The problem remains if we use the HP-definition – as HH does – instead of our working definition.

We can set this straight by looking instead at $P_{T^{Normal(b)*}}(\neg E | do(\neg C))$, so that we re-establish the counterfactual nature of our definition. (As $T^{**}|do(\neg C) = T^{*}|do(\neg C)$, this is equivalent to $P_{(T^{*})^{Normal(b)}}(\neg E | do(\neg C))$, which no longer mentions the artificial theory T^{**}.) However, by making this move we no longer take into account the (ab)normality of C itself, whereas research shows extensively that causal judgments regarding an event are often influenced by how normal it was (Kahneman, 1986; Knobe and Fraser, 2008; Hitchcock and Knobe, 2009). (This effect is not limited to normative contexts. For example, the lighting of a match is usually judged a cause of a fire, whereas the presence of oxygen is considered so normal that it isn't.) Hence we should factor in this normality, which is expressed by $P_{T^{Normal(b)}}(\neg C)$.

Definition 12. *[First refinement of Definition 11] Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an actual cause of E in (T, b) iff $P_{(T^{*})^{Normal(b)}}(\neg E | do(\neg C)) * P_{T^{Normal(b)}}(\neg C) > 0$.*

As the following theorem shows, our new choice only makes a difference in a limited set of cases.

Theorem 8. *If $R(C)$ contains a non-deterministic law or $P_{T^{Normal(b)}}(\neg C) = 0$, then*

$$P_{T^{Normal(b)*}}(\neg E \wedge \neg C) = P_{(T^{*})^{Normal(b)}}(\neg E | do(\neg C)) * P_{T^{Normal(b)}}(\neg C)$$

If there is a deterministic $r \in R(C)$ and $P_{T^{Normal(b)}}(\neg C) > 0$, as in the example shown, then contrary to the left-hand side of the equation, the proposed adjustment on the right-hand side of the equation gives the desired result $1 * 0.9 = 0.9$.

8. The Importance of Probabilities

Because the HH-approach lacks the quantification of normality offered by probabilities, they dismiss entirely all witnesses that are less normal than the actual world. A direct consequence is that any typical event – i.e., $P > 0.5$ – is never a cause, which is quite radical. By using probabilities, this qualitative criterion is no longer necessary: less normal witnesses simply influence our causal judgment less. Further, HH order causes solely by looking at the best witnesses. We now present an example which illustrates the benefit of both abandoning their criterion, and aggregating the normality of witnesses to order causes, without sacrificing the influence of normality.

Imagine you enter a contest. If a 10-sided die lands 1, you win a car. If not, you get a 100 more throws. If all of them land higher than 1, then you also win the car. The first throw lands 1, and you win the car.

It's hard to imagine anyone objecting to the judgment that the first throw is a cause of you winning the car. Yet that is exactly what we get when applying either Definition 11 or the improved Definition 12. The following theory T describes the set-up of the contest, where $Throw(i,j)$ means that the i-th throw landed j or smaller.

$$
\begin{aligned}
&(Throw(1,1):0.1) \leftarrow .\\
&(Throw(2,1):0.1) \leftarrow \neg Throw(1,1).\\
&(Throw(3,1):0.1) \leftarrow \neg Throw(1,1) \wedge \neg Throw(2,1).\\
&\qquad\qquad ...\\
&\qquad\qquad WinCar \leftarrow Throw(1,1).\\
&\qquad\qquad WinCar \leftarrow \neg Throw(2,1) \wedge ... \wedge \neg Throw(100,1).
\end{aligned}
$$

The normal refinement of T according to the story is given by:

$$
\begin{aligned}
&(Throw(1,1):0.1) \leftarrow .\\
&\qquad\qquad WinCar \leftarrow Throw(1,1).\\
&\qquad\qquad WinCar \leftarrow \neg Throw(2,1) \wedge ... \wedge \neg Throw(100,1).
\end{aligned}
$$

We get that $P_{(T^*)^{Normal(b)}}(\neg WinCar | do(\neg Throws(1,1))) = 0$, and thus $Throws(1,1)$ is not a cause of $WinCar$. In terms of HH: although $\neg Throw(1,1) \wedge \neg Throw(2,1) \wedge ... \wedge \neg Throw(100,1) \wedge WinCar$, is very unlikely, it is the only candidate witness. To see why, recall that a witness needs to have $\neg Throws(1,1)$, and should be at least as normal as the actual world. In every other world with $\neg Throws(1,1)$, at least one of the $Throws(i,1)$ is true, and hence it is less normal. But in a witness it should hold that $\neg WinCar$, so there is no witness for $Throws(1,1)$ being a cause of $WinCar$.

We can fix this problem by considering the theory $(T^*)^{NN}$ instead of $(T^*)^{Normal(b)}$. This leads us to another refinement of our original definition:

Definition 13. *[Second refinement of Definition 11] Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an actual cause of E in (T,b) iff $P_{(T^*)^{NN}}(\neg E | do(\neg C)) * P_{T^{Normal(b)}}(\neg C) > 0$.*

The theory $(T^*)^{NN}$ in this case is simply equal to T, but for the first law being $Throw(1,1) \leftarrow$. Hence the probability of not winning the car given that the first throw does not land 1 is pretty much 1, and the value in the equation becomes approximately 0.9, indicating $Throw(1,1)$ to be a very strong cause of $WinCar$.

Note that we only obtain this high value because the probability $P_{(T^*)^{NN}}$ aggregates the probabilities of all witnesses. If we would instead follow HH in considering only the best witness (in this case the story with $\neg Throw(1,1) \wedge Throw(2,1) \wedge \neg Throw(3,1) \wedge ... \wedge \neg Throw(100,1)$), we would obtain the much lower and less intuitive probability of 0.09.

Now imagine the same story, with a slight variation to the rules of the contest: you win the car on the first throw if the die lands anything under 7. Hence the first head changes to $Throw(1,6):0.6$, making it a typical outcome. Therefore the first law becomes deterministic in $T^{PN(b)}$, giving that $P_{T^{Normal(b)}}(\neg Throw(1,6)) = 0$, which again results in the counterintuitive judgment that the first throw in no way caused you to win the car.

We therefore suggest to use T^{NN} in the second factor of the inequality rather than $T^{Normal(b)}$, making use of the gradual measurement offered by probabilities. Applying this idea to the example, we get the result that $Throw(1,6)$ has causal strength 0.4. This value is smaller than before, because the cause is now less atypical.

8.1. The final definition

This brings us to our final extension to a definition of actual causation.

Definition 14 (Extension of actual causation)**.** *Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an actual cause of E in (T,b) if and only if $P_{(T^*)^{NN}}(\neg E | do(\neg C)) * P_{T^{NN}}(\neg C) > 0$.*

9. Conclusion and Related Work

In this paper we have used the formal language of CP-logic to formulate a general definition of actual causation, which we used to express four specific definitions: a previous proposal of our own, and three definitions based on the work of Hall. By moving from the deterministic context of neuron diagrams to the non-deterministic context of CP-logic, the latter definitions improve on the original ones in two ways: they can deal with a wider class of examples, and they allow for a graded judgment of actual causation in the form of a conditional probability. Also, comparison between the definitions is facilitated by presenting them as various ways of filling in two central concepts.

As mentioned, many other definitions exist in the counterfactual tradition. Rather than arguing for or against a particular definition, our aim in the first part of this work was to develop a general parametrised definition as a tool for constructing, comparing, and modifying different definitions in a systematic way. We briefly discuss some other definitions in order to understand the relation of our work to that of others.

The most influential definition of actual causation to date is the *HP definition* of Halpern and Pearl (2005a). This definition is expressed using structural equations as they are developed by Pearl (2000). Despite – or because of – its popularity, it has been subjected to much criticism. Restricting ourselves to authors working within the counterfactual tradition in the spirit of Lewis (1973), we can divide the criticism into two types.

The first type criticises not just the HP definition itself, but also its formulation in terms of structural equations. We already mentioned Hall (2007) as a prominent example. In previous work we have also criticised the HP definition and structural equations in general (Beckers and Vennekens, 2012; Vennekens, 2011). The current paper is a continuation of both lines of work. While structural equations are useful for a variety of purposes, we feel they lack certain key features when it comes to actual causation, which are present in CP-logic: true non-determinism in the endogenous part of the model, the distinction between default and deviant values, and a temporal semantics. It is possible to extend the language of structural models in various ways to incorporate such features (see, e.g., (Halpern and Pearl, 2005a; Hitchcock, 2007; Fenton-Glynn, 2015)), or to change the representation of specific examples in such a way that the need for them is avoided. Nevertheless, we feel that the fact that all of these features are integrated into CP-logic in a natural way makes the latter a suitable language for the study of actual causation.

In this paper we have exploited this fact by developing a general framework for actual causation in the context of CP-logic. The essence of this framework lies in the concepts of *Intrinsicness* and *Irrelevance* on the one hand, and the extension to actual causation on the other (Definitions 2 and 14). While similar concepts could be defined in the context of structural models, their definition is much more straightforward in the temporal, non-deteministic semantics of CP-logic.

The second type of criticism claims that although the HP definition is on the right track, it requires some adjustments in order to handle certain convincing counterexamples. As a result, several authors have proposed definitions using structural equations that are variants of the HP definition (Woodward, 2003; Hitchcock, 2001, 2007; Weslake, 2015; Halpern, 2015). Weslake (2015) offers an insightful comparison of several of these variants, concluding that none of them succeed in dealing with all counterexamples in a satisfactory manner. Although we have refrained from discussing many examples, both the BV12 and the Hall07 definition fare better on most of these problematic examples than the definitions mentioned. For completeness, we should also mention that Hitchcock (2009) proposes several counterexamples to Hall's definition.

The basic idea behind the HP definition and its variants formulated using structural equations is very similar to the idea behind our general definition formulated using CP-logic: construct variations of the causal model using information from the actual story, and check if there is counterfactual dependence of the effect on the candidate cause in one of these variations. Besides the use of different formal languages, the difference between both approaches is twofold. First, we use probabilities to quantify the importance of causes. This proved particularly helpful in our discussion of an extension to actual causation. The second, more fundamental, difference lies in the methodology of constructing variations of a causal model.

The HP-like definitions construct variations using so-called *structural contingencies*. A structural contingency is some set of interventions on a structural model. Different approaches differ in which structural contingencies they allow. Typically, no principled account is given of why certain structural contingencies should be allowed or not. Instead, this is decided in an *ad hoc* manner, based on whether allowing them provides the right answer for certain problematic examples. Comparisons between different approach are therefore typically also reduced to a tally of (in-)correctly handled examples.

As we have seen, in our approach on the other hand, the construction of variations is determined by the *Intrinsicness* and *Irrelevance* functions. Therefore different instantiations of our general definition can be compared directly, and can be defended by means of principled arguments in favour of particular definitions for these functions. Given the overwhelming amount of problematic examples and the conflicting intuitions that come with them, we believe a systematic approach to defining actual causation is the right way forward.

In the second part of this work, we showed how our general definition lends itself to an extension of actual causation that continues upon the work of Halpern and Hitchcock (2015). It incorporates the main points raised by Halpern and Hitchcock: (1) it allows normative considerations and (2) is able to factor in the normality of the cause given the context. This extension is useful in normative disciplines, such as law and ethics, and takes into account the context sensitivity of causal judgements suggested by recent findings in experimental psychology (Knobe and Fraser, 2008; Moore, 2009; Hitchcock and Knobe, 2009). Our account improves on the HH account in several ways:

- By using CP-logic our account can also be applied to non-deterministic examples.
- Separating normative from statistical normality allows for a more accurate description of the domain.
- Since we no longer refer to the actual world in the second factor, we can use strict norms.
- The reader may verify that our approach is able to deal with all of the examples given by Halpern and Hitchcock (2015) equally well as Definition 11.
- It can also properly handle the examples from Sections 7 and 8, as opposed to Definition 11.

Appendix A.

To facilitate the proof of the first theorem, we introduce the following lemma.

Lemma 1. *Given a neuron diagram D with its corresponding equations M, and an assignment to its variables V. Consider the CP-logic theory T, and a story b, that we get when applying the translation from Section 3.3.1. Then a neuron diagram R is a reduction of D in which both C and E occur iff its translation d – another branch of T – is a (C,E)-reduction of b.*

Proof. Assume we have a reduction R of a neuron diagram D, and b is the story corresponding to D. As R is simply a different assignment to the variables occurring in D, brought about by the same equations that existed for D, this reduction corresponds to another branch d of T, in which C and E hold in its leaf. Moreover, R can be constructed starting from D by changing some of the exogenous variables, say $\mathbf{U}'$, from their actual values to their default value, and then updating the endogenous variables in accordance with the deterministic equations. It being a reduction, this caused no new variables to take on their deviant value in comparison to D. Let r be a law that occurs in d.

If r is non-deterministic, it must be one of the laws representing a variable V that is determined directly by the exogenous variables, i.e., a law with an empty body, and hence it was also applied in b. R being a reduction, either V has the same value in R as in the original diagram, or it has its default value. In the former case, this means that $r_d = r_b$, in the latter case $r_d = 0$, both of which satisfy the requirement for d being a reduction.

If r is deterministic, the precondition for r has to be fulfilled in d, causing some variable V to take on its deviant value. The same must hold true of the precondition for the equation for V, and thus V takes on its deviant value in R as well, implying it did so in D too. Therefore there must have been some law applied in b that made V take on its deviant value as well. From this it follows that d is a (C,E)-reduction of b.

Now assume we have a theory T and a story b that form the translation of a neuron diagram D and the states of its neurons, such that C and E hold in b, and that d is a (C,E)-reduction of b. As the leaf of d contains an assignment to all of the variables that satisfies the equations of M, there is a neuron diagram R that corresponds to d. We can easily go over all the previous steps in the other direction, to conclude that R is a reduction of D in which C and E are true. □

Theorem 3. *Given a neuron diagram with its corresponding equations M, and an assignment to its variables $\mathbf{V}$. Consider the CP-logic theory T and story b that we get when applying the translation from Section 3.3.1. Then C is an actual cause of E in the diagram according to Hall's definition iff C is an actual cause of E in b and T according to Definition 3.*

Proof. We start with the implication from left to right. Assume we have a neuron diagram D, in which both C and E fire. This translates into a theory T and a story b, for which C and E hold in its leaf. Further, assume there is a reduction R of this diagram, in which both C and E continue to hold, and in this reduction, if $C = 0$; then $E = 0$. By the above lemma, this translates into a (C,E)-reduction of b, say d.

In R, if $C = 0$; then $E = 0$. The conditional $C = 0$ is interpreted as a counterfactual locution, and corresponds to $do(\neg C)$. As there are no non-deterministic laws with non-empty preconditions, T^d is simply the deterministic theory that determines the same assignment as R, meaning $P_{T^d}(\neg E|do(\neg C)) = 1$, which concludes this part of the proof.

Now assume we have a theory T and a story b that form the translation of a neuron diagram D, such that C and E hold in b, and that d is a (C,E)-reduction of b for which the given inequality holds. By the above lemma, the translation of d, say R, is reduction of D in which C and E occur. As mentioned in the previous paragraph, T^d simply corresponds to an assignment of values to the variables occurring in D that follows its equations. Since R describes this same assignment, in R too if $C = 0$; then $E = 0$. This concludes the proof. □

Theorem 4. *If* $(\exists d \in Red_b^{(C,E)} : P_{T^d}(\neg E|do(\neg C)) > 0)$ *then* $P_{T^{Nec(b)}}(\neg E|do(\neg C)) > 0$. *If b is simple, then the reverse implication holds as well.*

Proof. We start with proving the first implication. Assume we have a $d \in Red_b^{(C,E)}$ such that $P_{T^d}(\neg E|do(\neg C)) > 0$. This implies that there is at least one branch e of a probability tree of $T^d|do(\neg C)$ for which $\neg E$ holds in its leaf. We prove by induction on the length of e that this implies the existence of a similar branch e' of a probability tree of $T^{Nec(b)}|do(\neg C)$ for which $\neg E$ holds in its leaf, which is what is required to establish the theorem.

Base case: if e consists of a single node – i.e., the root node where all atoms are false – then this means that no laws of $T^d|do(\neg C)$ can be applied. Since the bodies of the laws in $T^{Nec(b)}|do(\neg C)$ are identical to those of the laws in $T^d|do(\neg C)$, we simply have $e' = e$.

Induction case: Assume we have a sub-branch e_n of e with length $n > 1$, starting from the root node, and that we also have a structurally identical sub-branch e'_n. By it being structurally identical we mean that they are identical except for the fact that they may have different probabilities along the edges.

If $e_n = e$, then no more laws can be applied in the final node of e_n. This must then hold for the final node of e'_n as well, so we are finished. Otherwise, we know that there is a sub-branch e_{n+1} which extends e_n along e with a node O. Assume that the law which was applied to get to O is r.

If r is deterministic, then r occurs in $T^d|do(\neg C)$ exactly as it does in $T^{Nec(b)}|do(\neg C)$. Since both branches are structurally identical, e'_n can be extended in the exact same manner as e_n, so there has to be a probability tree of $T^{Nec(b)}|do(\neg C)$ in which there is a sub-branch e'_{n+1} with the desired properties. So assume r is non-deterministic.

First assume $r \notin Laws_d$. This implies that $r \notin Nec(b)$. So as in the deterministic case, r occurs in $T^d|do(\neg C)$ exactly as it does in $T^{Nec(b)}|do(\neg C)$, and the branch can be extended in the same manner.

Now assume $r \in Laws_d$. If also $r \in Nec(b)$, we know that $r_d = r_b = r_{Nec}$ and hence the previous argument holds. Remains the possibility that $r \notin Nec(b)$. As in the deterministic case, because r can be applied in the final node of e'_n there has to be a probability tree of $T^{Nec(b)}|do(\neg C)$ with a sub-branch like e'_n where r is applied next.

Assume $r_d = A$. Since A was the outcome of r in d, the law r as it appears in T – and also in $T^{Nec(b)}|do(\neg C)$ – contains A in its head with some probability attached to it. Therefore the final node of e'_n in the said probability tree has one child-node which contains A, extending e'_n into a sub-branch e'_{n+1} with the desired properties. This concludes this part of the proof.

Now we prove that if b is simple, the reverse implication holds as well.

Assume $P_{T^{Nec(b)}}(\neg E|do(\neg C)) > 0$. This implies that there is at least one branch e of a probability tree of $T^{Nec(b)}|do(\neg C)$ for which $\neg E$ holds in its leaf. We can repeat the first steps of the previous implication, so that we again arrive at a law r which was applied to get to a node O.

The branch e' we are considering occurs in a probability tree of a (C,E)-reduction, say f. First assume $r \in Nec(b)$. By definition, this implies that also $r \in Laws_f \wedge r_{Nec} = r_f$, and we can apply the reasoning from above. Likewise as above, we can apply this reasoning to all other cases, except the one where $r \notin Nec(b)$, r is non-deterministic, and $r \in Laws_f$. Assume the law r has effect A in the branch e we are considering. If $r_f = A$, then we are back to our familiar situation, so therefore assume $r_f = B$, and $A \neq B$.

Since b is simple, A and B are the only two possible effects of r. Further, remark that $r \in Laws_b \setminus Nec(b)$. This implies the existence of a (C,E)-reduction g that is identical to f up to the application of r, but such that $r_g \neq r_f$, and

thus $r_g = A = r_e$ meaning there is a branch in a probability tree of g that is structurally identical to e up to O. This concludes the proof of the theorem. □

Theorem 5. *If b is simple, then a non-deterministic law r is necessary iff none of the r-variants of b is a (C,E)-reduction.*

Proof. We denote the node in b in which r is applied by N. This node has two children, one representing the selection of r_b, and its sibling, say O. Thus, the r-variants of b are all branches passing through O.

We start with the implication from left to right, so we assume r is necessary. Assume $r_b = A$, then there is no $d \in Red_b^{(C,E)}$ for which $r_d \neq A$, hence there is no (C,E)-reduction which passes through O.

Remains the implication from right to left. Assume we have a law r such that there is no (C,E)-reduction passing through O. We proceed with a reductio ad absurdum, so we assume r is not necessary.

Clearly b is a (C,E)-reduction of itself, and also $r \in Laws_b \setminus Nec(b)$. Hence, by b's simplicity, there is a (C,E)-reduction e which is identical to b up to the application of r, but for which $r_e \neq r_b$. Thus e passes through O, contradicting the assumption that r is necessary. This concludes the proof. □

Theorem 6. *Given a neuron diagram with its corresponding equations M, and an assignment to its variables* **V**. *Consider the CP-logic theory T, and a story b, that we get when applying the translation from Section 3.3.1. C is a producer of E in the diagram according to Hall iff C is an actual cause of E given* Production-Irr *and* Production-Intr.

Proof. First we need to explain some terminology that Hall uses. A *structure* is a temporal sequence of sets of events, which unfold according to the equations of some neuron diagram. A branch, or a sub-branch, would be the corresponding concept in CP-logic.

Two structures are said to *match intrinsically* when they are represented in an identical manner. The reason why Hall uses this term, is because even though we use the same variable for an event occurring in different circumstances, strictly speaking they are not the same. This is mainly an ontological issue, which need not detain us for our present purposes.

A set of events S is said to be *sufficient* for another event E, if the fact that E occurs follows from the causal laws, together with the premisse that S occurs at some time t, and no other events occur at this time. A set is *minimally* sufficient if it is sufficient, and no proper subset is. To understand this, note that the ambiguity of the relation between an event and the value of a variable that we noted earlier, resurfaces here. In the context of neuron diagrams, events are temporal, and occur during the time-period that a neuron fires, i.e, becomes true. However, at any later time-point, the variable corresponding to this neuron will remain to be true, implying that the value of the variable has shifted in meaning from "the neuron fires" to "the neuron has fired". Given this interpretation, it is natural to translate Hall's notion of an event into CP-logic as the application of a law, making a variable true, as we have done.

A further detail to be cleared out, is that in the context of neuron diagrams there can be simultaneous events, since multiple neurons can fire at the same time. In CP-logic, in each node only one law is allowed to be applied, hence this translates to two consecutive edges in a branch. Therefore it is not the case that each node-edge pair in a branch corresponds to a separate time-point, but rather sets of consecutive pairs – with variable size – do. Given such a set, then for each variable that was the result of the application of a law belonging to it, it holds that its corresponding event occurs at the next time-point, corresponding to the next set of nodes further down the branch. All the variables occuring in the bodies of the laws in this set, represent events that occur during this time-point.

Now we can state the precise definition of production as it occurs in (Hall, 2004, p.25).

> We begin as before, by supposing that E occurs at t', and that t is an earlier time such that at each time between t and t', there is a unique minimally sufficient set for E. But now we add the requirement that whenever t_0 and t_1 are two such times ($t_0 < t_1$) and S_0 and S_1 the corresponding minimally sufficient sets, then
>
> - for each element of S_1, there is at t_0 a unique minimally sufficient set; and
> - the union of these minimally sufficient sets is S_0.

...

> Given some event E occurring at time t' and given some earlier time t, we will say that E has a *p*ure causal history back to time t just in case there is, at every time between t and t', a unique minimally sufficient set for E, and the collection of these sets meets the two foregoing constraints. We will call the structure consisting of the members of these sets the "pure causal history" of E, back to time t. We will say that C is a proximate cause of E just in case C and E belong to some structure of events S for which there is at least one nomologically possible structure S' such that (i) S' intrinsically matches S; and (ii) S' consists of an E-duplicate, together with a pure causal history of this E-duplicate back to some earlier time. (In easy cases, S will itself be the needed duplicate structure.) Production, finally, is defined as the ancestral [i.e., the transitive closure] of proximate causation.

We will start with the implication from left to right. So assume we have a neuron diagram D, in which C is a producer of E. Say T is the CP-logic theory that is the translation of the equations of the diagram, and b is the branch representing the story. We already know that C and E hold in the leaf of b. We need to proof that $P_{T^*}(\neg E|do(\neg C)) > 0$. The theory T^* only contains deterministic laws, and no disjunctions, hence all its laws are of the form: $V \leftarrow A \wedge A' \wedge ... \wedge \neg B \wedge \neg B'$, where the number of positive literals in the conjunction is at least one. Therefore any probability tree for T^* consists out of only one branch, determining a unique assignment for all the variables. Further, even though the theory T may contain several laws in which a variable occurs in the head, because of our irrelevance criterion T^* contains exactly one law for every variable that is true. So for every true variable in this assignment, there is a unique chain of laws – neglecting the order – which needs to be applied to make this variable true. For any such variable V, we will say that it depends on all of the variables occurring positively in the body of a law in this chain. Clearly, if any true variable changes its value in this assignment, then all variables which depend on it become false.

As a first case, assume C is a proximate cause of E. We start by assuming that circumstances are nice, meaning that D contains itself a structure S which is a pure causal history of E. This means that in the actual story b, C is part of a unique minimally sufficient set for E. From this it follows that in T^*, C figures positively in one of the laws on which E depends. Hence, if we apply $do(\neg C)$, then E will no longer hold.

Now assume that there is a structure S occurring in D, such that there exists another diagram, say D', in which this structure occurs as well, and forms a pure causal history of E. This diagram corresponds to a branch of T, say d, That means that in T^*_d – i.e., the theory T^* constructed out $d - C$ occurs positively in the unique chain of laws which can make E true. But as all events in S also occur in D, at the same moments as they do in D', that means that C must also occur positively in the unique chain of laws for E in the theory T^*_b. Hence, E depends on C in the theory T^*_b as well.

Now look at the more general case, in which C occurs in a chain of proximate causes, that leads up to E. I.e, in D, C is the proximate cause of some variable V_1, which in turn is the proximate cause of some variable V_2, and so on until we get to E. We know from the previous discussion, that this implies in T^* that $do(\neg C)$ then $\neg V_1$, and $do(\neg V_1)$ then $\neg V_2$, and so on. Given what we know about T^*, it directly follows that when we apply $do(\neg C)$, then $\neg E$. This concludes this part of the proof.

We continue with the implication from right to left. So assume that we are given again a neuron diagram and a corresponding story b, and that we know $P_{T^*}(\neg E|do(\neg C)) > 0$. From our earlier analysis of T^*, we know that this means that C occurs positively in the unique chain of laws that can make E true according to T^*. From this chain of laws, we start from the one causing E and from there pick out a series that gets us to a law where C occurs positively in the body. More concretely, we take a series of the form: $E \leftarrow ...A \wedge ...$, $A \leftarrow ...D \wedge ...$, and so on until we get at a law $Z \leftarrow ...C \wedge ....$ By definition of production, it suffices to prove that in this chain, each of the variables in the body is a proximate cause of the variable in the head.

Take such a law $V \leftarrow ...W \wedge ....$ At the time that this law is applied, W clearly is a member of a sufficient set of events for V, which occurs at the next time point. Say S_0 is the set of all events that occur together with W that figure in the body of this law, and S_1 is the set $\{V\}$ that occurs at the next time-point, then the structure consisting precisely of S_0 and S_1 and nothing else forms a pure causal history of V containing W. The same reasoning applies to all laws of the chain. This concludes the proof. □

Appendix B.

To facilitate the proof of Theorem 7, we introduce the following lemma.

Lemma 2. *Given an extended model and exogenous assignment $(M,\succeq,\mathbf{u})$, and a theory and branch (T,b) that are derived from $(M,\succeq,\mathbf{u})$ in the sense described in Section 6. Then for any world w, and a branch d of a probability tree from T that corresponds to it, it holds that $w \succeq s_{\mathbf{u}}$ iff d occurs in a probability tree of $T^{Normal(b)}$.*

Proof. We know that b is a branch in a probability tree from T such that $Leaf_b$ has the same assignment as $s_{\mathbf{u}}$. Recall that T consists of two categories of laws. First there are those corresponding to the equations for the endogenous variables which depend on other endogenous variables, which are deterministic and thus re-appear in $T^{Normal(b)}$ unchanged. Second there are those corresponding to the endogenous variables which directly depend on the exogenous variables, which take the form $X : p\{q\} \leftarrow$, where the second probability need not be present.

Assume we have a world w such that $w \succeq s_{\mathbf{u}}$. Any world that satisfies the equations of M follows deterministically from an assignment to all exogenous variables. As $s_{\mathbf{u}}$ is a world that satisfies the equations, and w is at least as normal, it also satisfies the equations. Hence there is an exogenous assignment $\mathbf{u}'$ which determines w. In CP-logic, such an assignment corresponds to choosing particular disjuncts in the heads of all laws from the second category.

Concretely, this means that for each law/equation of the second category, the value of the corresponding variable X is at least as typical in $w = s_{\mathbf{u}'}$ as it is in $s_{\mathbf{u}}$. Denote by X_w and X_s the values X takes in the worlds w and $s_{\mathbf{u}}$ respectively. By construction of $T^{Normal(b)}$, the disjuncts which are at least typical as X_s – be it in the statistical or in the normative sense – still appear in the law for X in $T^{Normal(b)}$, and hence can be chosen when this law is applied. Therefore the branches corresponding to w from the probability trees of T also appear in the probability trees of $T^{Normal(b)}$, be it that the values of the probabilities may have changed.

Now assume we have a branch d corresponding to a world w, that occurs in a probability tree of $T^{Normal(b)}$. We can simply reverse the correspondence between the choices of disjuncts and an exogenous assignment, to obtain that $w \succeq s_{\mathbf{u}}$.

□

Theorem 7. *C is an HH-actual cause of E in an extended model and exogenous assignment $(M,\succeq,\mathbf{u})$ iff C is an HH-CP-logic-actual cause of E in (T,b), where (T,b) is derived from $(M,\succeq,\mathbf{u})$ in the sense described in Section 6.*

Proof. We begin with the implication from left to right. So assume we have an extended model and exogenous assignment $(M,\succeq,\mathbf{u})$, such that C and E hold in $s_{\mathbf{u}}$, and there is at least one witness w of C being an actual cause of E in $(M,\mathbf{u})$ such that $w \succeq s_{\mathbf{u}}$.

Recall that we assume the definition of actual causation at hand can be translated from structural models into an instantiation of our general definition. So we get that C is an actual cause of E in (T,b), and more specifically that any branch d that corresponds to w is a witness of this. Thus d appears in a probability tree of $T^*|do(\neg C)$.

By Lemma 2, we know that such a branch d also appears in a probability tree of $T^{Normal(b)}$.

We look separately at the two options regarding $R(C)$. First we assume that C is determined directly by the exogenous variables, meaning that $R(C)$ consists of a single non-deterministic law, say $r(C)$. Since d occurs in a tree of $T^{Normal(b)}$, and $\neg C$ holds in it, the empty disjunct remains present in the normal refinement of $r(C)$. By definition, T^{**} is simply $T^*|do(\neg C)$ with the normal refinement of $r(C)$. Therefore d also occurs in a tree of T^{**}.

Second, assume the laws in $R(C)$ are deterministic. Since d occurs in a tree of $T^{Normal(b)}$, which obviously contains C in the head of any law $r(C) \in R(C)$, the body for $r(C)$ cannot be satisfied in d. Thus all laws $R(C)$ are irrelevant to d. Since T^{**} and $T^*|do(\neg C)$ are identical but for $R(C)$, we can again conclude that d also occurs in a tree of T^{**}.

So in all cases we have that d occurs both in a tree of $T^{Normal(b)}$, and in a tree of T^{**}. This implies that the disjuncts chosen in the laws applied in d occur in the versions these laws take in both of these theories, with possibly different but strictly positive probabilities. Note that every law from $T^{Normal(b)*}$ either takes the form it has in T^{**} or it takes the form it has in $T^{Normal(b)}$. Therefore d also appears in $T^{Normal(b)*}$. It being a witness, $\neg C$ and $\neg E$ hold in it, and thus the stated probability is strictly positive.

Now we continue with the reverse implication. Assume we have an extended theory T, a story b such that C and E hold in it, and $P_{T^{Normal(b)*}}(\neg E \wedge \neg C) > 0$. This implies the existence of a branch d in $T^{Normal(b)*}$ such that both $\neg C$ and $\neg E$ holds.

Say r is a law from $T^{Normal(b)*}$. If r is intrinsic and $r \ni nR(C)$, it is deterministic, containing the single (possibly empty) disjunct r_d with associated probability 1. As r_d was the actual choice from b, by construction r_d also appears in the normal refinement of r, although the probability may be different. However, as long as we do not have strict norms, i.e., norms where p or q is 1, this probability will be strictly positive. A strict norm means that a violation of it is considered entirely abnormal, analogous to the occurrence of an event with zero probability. Since HH treat norms identical to statistical normality, and since the actual world was possible, it follows that the actual world is not entirely abnormal. Hence even if r_d was a violation of a norm, it will not have been a strict norm. (Our final definition from Section 8.1 does allow for strict norms.) Thus, we conclude that r_d occurs in the head of the versions of the law r we find in both T^* and $T^{Normal(b)}$. Because $r \not\ni R(C)$, we can say the same about $T^*|do(\neg C)$.

If r is not intrinsic and $r \notin R(C)$, it contains all of its original disjuncts when it occurs in T^*. Therefore it takes the same form in $T^{Normal(b)*}$ as it does in $T^{Normal(b)}$. Again we conclude that r_d occurs in the head of the versions of the law r we find in each of T^*, $T^{Normal(b)}$ and $T^*|do(\neg C)$.

This leaves us to consider the laws in $R(C)$. By definition, $T^{Normal(b)*}$ contains the same version of these laws as $T^{Normal(b)}$. From this and the previous paragraphs we can already conclude that any branch occuring in a tree of $T^{Normal(b)*}$ also occurs in a tree of $T^{Normal(b)}$. More specifically this holds for d. Thus by Lemma 2, it holds for the corresponding world w that $w \succeq s_u$.

If the body for each $r \in R(C)$ is false in d, then the precise form of the head of r is irrelevant for d. As the head of each $r \in R(C)$ is the only difference between T^{**} and $T^*|do(\neg C)$, we can again conclude that d also occurs in $T^*|do(\neg C)$.

Leaves us to consider the case that there is some $r \in R(C)$ for which the body is true in d. From the fact that d – in which $\neg C$ holds – occurs in $T^{Normal(b)}$, we can infer that r is a non-deterministic law, and thus the only member of $R(C)$. Taken together with the knowledge that the disjunct containing C was chosen in b, it follows that the normal refinement of r contains both C and the empty disjunct in its head. Furthermore, in d the empty disjunct was chosen. These observations taken together imply that the disjunct of r chosen in d occurs in the head of the versions of r we find in both T^{**} and $T^*|do(\neg C)$. Once more we conclude that d also occurs in $T^*|do(\neg C)$.

Thus d is a witness for C being an actual cause of E in (T,b). Therefore the world w corresponding to d is a witness for C being an actual cause of E in (M,u). Together with the fact that $w \succeq s_u$, the conclusion follows.

□

Theorem 8. *If $R(C)$ contains a non-deterministic law or $P_{T^{Normal(b)}}(\neg C) = 0$, then*

$$P_{T^{Normal(b)*}}(\neg E \wedge \neg C) = P_{(T^*)^{Normal(b)}}(\neg E|do(\neg C)) * P_{T^{Normal(b)}}(\neg C)$$

Proof. First we examine the case where $P_{T^{Normal(b)}}(\neg C) = 0$. This implies that the right-hand side of the equation is 0. Also, any branch from a tree $T^{Normal(b)*}$ occurs as well in a tree of $T^{Normal(b)}$, so $P_{T^{Normal(b)*}}(\neg C) = 0$ and the left-hand side is also equal to 0.

This leaves us to consider the case where $P_{T^{Normal(b)}}(\neg C) > 0$ and the unique $r(C) \in R(C)$ is non-deterministic.

In this case $P_{T^{Normal(b)*}}(\neg C) = P_{T^{Normal(b)}}(\neg C)$, so we have:

$$P_{T^{Normal(b)*}}(\neg E \wedge \neg C) = P_{T^{Normal(b)*}}(\neg E \wedge \neg C) * P_{T^{Normal(b)}}(\neg C)/P_{T^{Normal(b)*}}(\neg C) = P_{T^{Normal(b)*}}(\neg E|\neg C) * P_{T^{Normal(b)}}(\neg C)$$

Further, conditioning on $\neg C$ when C only occurs in a vacuous non-deterministic law, is identical to looking at the intervention $do(\neg C)$, thus the list of equalities continues:

$= P_{T^{Normal(b)*}}(\neg E|do(\neg C)) * P_{T^{Normal(b)}}(\neg C)$. Also, $T^{Normal(b)*}|do(\neg C) = (T^*)^{Normal(b)*}|do(\neg C)$, which brings us to the desired conclusion.

□

Acknowledgements

Sander Beckers was funded by a Ph.D. grant of the Flemish Agency for Innovation by Science and Technology (IWT-Vlaanderen).

References

Beckers, S., Vennekens, J., 2012. Counterfactual dependency and actual causation in cp-logic and structural models: a comparison. In: Proceedings of the Sixth STAIRS. pp. 35–46.

Beckers, S., Vennekens, J., 2015a. Combining probabilistic, causal, and normative reasoning in cp-logic. In: 12th International Symposium on Logical Formalizations of Commonsense Reasoning. pp. 32–38.

Beckers, S., Vennekens, J., 2015b. Towards a general framework for actual causation using cp-logic. In: Proceedings of the 2nd International Workshop on Probabilistic Logic Programming co-located with ICLP. Vol. 1413. pp. 19–38.

Fenton-Glynn, L., 2015. A proposed probabilistic extension of the halpern and pearl definition of 'actual a proposed probabilistic extension of the halpern and pearl definition of 'actual cause'. British journal for the philosophy of science forthcoming.

Hall, N., 2004. Two concepts of causation. In: Collins, J., Hall, N., Paul, L. A. (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, L. A., 2003. Causation and Preemption. Oxford University Press.

Halpern, J., 2015. A modification of the halpern-pearl definition of causality. In: Proceedings of the 24th IJCAI. AAAI Press, pp. 3022–3033.

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

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

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.

Kahneman, Daniel; Miller, D. T., 1986. Norm theory: comparing reality to its alternatives. Psychological Review 94 (2), 136–153.

Knobe, J., Fraser, B., 2008. Causal judgment and moral judgment: Two experiments. In: Sinnott-Armstrong, W. (Ed.), Moral Psychology. MIT Press.

Lewis, D., 1973. Causation. Journal of Philosophy 70, 113–126.

Moore, M. S., 2009. Causation and Responsibility. OUP Oxford.

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

Sato, T., 1995. A statistical learning method for logic programs with distribution semantics. In: Proceedings of the 12th International Conference on Logic Programming. pp. 715–729.

Shafer, G., 1996. The Art of Causal Conjecture. Artificial Management. MIT Press.
URL `http://books.google.be/books?id=sY7os7OCykUC`

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

Vennekens, J., Denecker, M., Bruynooghe, M., 2009. CP-logic: A language of probabilistic causal laws and its relation to logic programming. Theory and Practice of Logic Programming 9, 245–308.

Vennekens, J., Denecker, M., Bruynooghe, M., 2010. Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In: JELIA. pp. 313–325.

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.