Combining Probabilistic, Causal, and Normative Reasoning in CP-Logic

Sander Beckers and **Joost Vennekens**

KU Leuven – University of Leuven

{Sander.Beckers,Joost.Vennekens}@cs.kuleuven.be

Abstract

In recent years the search for a proper formal definition of actual causation – i.e., the relation of cause-effect as it is instantiated in specific observations, rather than general causal relations – has taken on impressive proportions. In part this is due to the insight that this concept plays a fundamental role in many different fields, such as legal theory, engineering, medicine, ethics, etc. Because of this diversity in applications, some researchers have shifted focus from a single idealized definition towards a more pragmatic, context-based account. For instance, recent work by Halpern and Hitchcock draws on empirical research regarding people's causal judgments, to suggest a graded and context-sensitive notion of causation. Although we sympathize with many of their observations, their restriction to a merely qualitative ordering runs into trouble for more complex examples. Therefore we aim to improve on their approach, by using the formal language of CP-logic (Causal Probabilistic logic), and the framework for defining actual causation that was developed by the current authors using it. First we rephrase their ideas into our quantitative, probabilistic setting, after which we modify it to accommodate a greater class of examples. Further, we introduce a formal distinction between statistical and normative considerations.

1 Introduction

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 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 whether 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 two 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. In this paper we offer an alternative, in which statistical normality is represented 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. As a technical tool for our analysis, we will use a normative extension of a general framework for the study of actual causation (Beckers and Vennekens) that was defined in the language of CP-logic (Causal Probabilistic logic) (Vennekens, Denecker, and Bruynooghe 2009). An important property of our technical approach is that we implement the concept of normality by means of a syntactic transformation, similar to how previous work has defined the concept of interventions in CP-logic (Vennekens, Denecker, and Bruynooghe 2010). This allows us to improve on the HH approach, by properly accommodating the counterfactual nature of causation. The result will be a more generally applicable and yet simpler approach.

In Section 2 we shortly present the formal language of CP-logic and use it to formulate a definition of actual causation. The following section presents the extension to actual causation by HH. We translate their work into the CP-logic framework in Section 4. Section 5 contains a first improvement to this translation, followed by some examples and our final extension to actual causation in Section 6.

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:

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.

2 Preliminaries: CP-logic

We give a short, informal introduction to CP-logic. A detailed description can be found in (Vennekens, Denecker, and Bruynooghe 2009; 2010). The basic syntactical unit of CP-logic is a CP-law, which takes the general form of $Head \leftarrow Body$. The body can in general consist of any first-order logic formula. However, in this paper, we restrict our attention to grounded formulas in CNF. 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, we take this to mean that, 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. To illustrate this, we describe the relevant events on the Monday morning from our running example. 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:

$$(Prof : 0.7) \leftarrow . \tag{1}$$
$$(Assistant : 0.8) \leftarrow . \tag{2}$$
$$NoPens \leftarrow Prof \wedge Assistant . \tag{3}$$

The first two of these causal laws are *vacuous* (i.e, they will be applied in every story) and *non-deterministic*. The first one results in either the professor taking the pen, or nothing at all. Since it is stated that professors typically do take pens, we simply chose some probability above 0.5. Similarly, the second law describes the possible behaviour of the assistant. The last law is *deterministic*, i.e., it only has one possible outcome (where we leave implicit the probability 1).

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
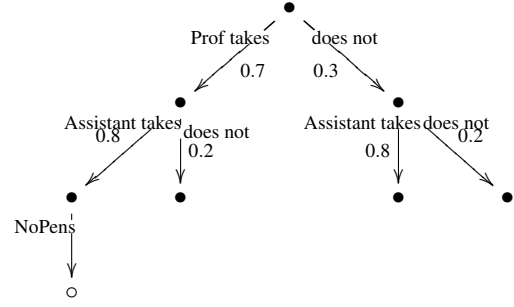

Figure 1: Probability tree for the Pen-vignette.

pens. The other stories consist in only the professor taking a pen, only the assistant doing so, or neither.

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 *Prof*, *Assistant* and *NoPens*. In the initial state of the domain (the root node), all atoms are assigned their *default* value **false**. In this example, initially there are still pens left, and neither the professor nor the assistant have taken one. 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. Because in this particular tree the professor arrives first, there are two child-nodes, corresponding to the law (1) being applied and resulting in a pen being taken (left child) or not (right child). In the former case, the assignment is updated by setting *Prof* to **true**, its *deviant* value. Similarly, the subsequent nodes represent the possible outcomes of the application of law (2), and whether or not this results in the lack of pens. The leftmost branch is thus the formal counterpart of the above story. A probability tree of a theory T defines an *a priori* probability distribution P_T over all stories 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 our example story is 0.56. We have shown here only one such probability tree, but we can construct another one 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 this 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.

2.1 Operations on CP-theories

We specify some operations on CP-logic theories that will be used throughout this paper. Assume we have a theory T and a branch b of one of T's probability trees, such that both C and E hold in its leaf. To make T deterministic *in accordance with the choices made in b*, means to transform T into T^b by replacing the heads of the laws that were applied in b with the disjuncts which were chosen from those heads in b.

For our example story b, T^b is:

$$Prof \leftarrow . \tag{4}$$

$$Assistant \leftarrow . \tag{5}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{6}$$

We will use Pearl's $do()$-operator to indicate an intervention (Pearl 2000). The intervention on a theory T that makes variable C 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 the professor from taking a pen, the resulting theory $T|do(\neg Prof)$ is given by:

$$\leftarrow . \tag{7}$$

$$(Assistant : 0.8) \leftarrow . \tag{8}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{9}$$

Laws with an empty head, such as (7), can also simply be omitted. The analogous operation $do(C)$ on a theory T corresponds to adding the deterministic law $C \leftarrow$.

2.2 Actual Causation in CP-logic

In recent years many proposals for defining actual causation have entered the scene, such as (Hall and Paul 2003; Hall 2004; 2007; Halpern and Pearl 2005; Vennekens 2011; Beckers and Vennekens 2012). In (Beckers and Vennekens) we developed a general definition, which encompasses several of these. The idea behind our general definition is that one should first modify the theory T into a theory T^* that takes into account the influence of the actual story b, and then judge causation by using the probabilistic extension of counterfactual dependence. The available operations to construct T^* are to make laws deterministic in accordance with the choices made in b, and to remove laws. A specific definition is obtained by specifying which laws fall under the first category – called the *intrinsic* laws and noted as Int – and those which fall under the second – called the *irrelevant* laws and noted as Irr.[1] We then construct T^* as $T \setminus (Irr \cup Int) \cup Int^b$.

Definition 1. *[Actual causation] Given a 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* $P_{T^*}(\neg E|do(\neg C)) > 0$.

If desired, the degree to which different events actually caused some effect can then be further compared by comparing their associated probabilities.

HH use the HP-definition (Halpern and Pearl 2005) as their working definition to illustrate their extension to actual causation. However, they stress the generality of their

---

[1] These specifications are implicitly parametrized with regard to a theory, a story, and events C and E.

approach, and mention that one could for example apply it to Hall's definition from (Hall 2007). In (Beckers and Vennekens), we reformulated Hall's definition as an instantiation of the general definition here presented. We will choose this reformulation to work with in this paper, because this makes it easier to apply the current discussion to other instantiations.

By $Leaf_b$ we denote the set of true variables in the leaf of a branch b. For a law r that was applied in b, we call the node in b that results from its application $Node_r^b$. Our working definition states that no laws are irrelevant, and determines the intrinsic laws as follows:

Intrinsic. *A non-deterministic law r of T that was applied in b is intrinsic iff there is no branch d passing through a sibling of* $Node_r^b$ *such that* $\{C, E\} \subseteq Leaf_d \subseteq Leaf_b$.

If we look at the probability tree in Figure 1, we see that $E = NoPens$ only occurs in $Leaf_b$, and hence both the laws (1) and (2) are intrinsic. This gives $T^* = T^b$. As there is counterfactual dependency of $NoPens$ on both $Assistant$ and $Prof$, they are full causes. However it was shown in (Knobe and Fraser 2008) that most people do not judge the assistant to be a cause, or at least far less so than the professor is. This illustrates how our causal judgment can be influenced by normative considerations.

3 The HH Extension to Actual Causation

In this section we succinctly present the graded, context-dependent approach to actual causation from (Halpern and Hitchcock). As with much work on actual causation, HH frame their ideas using structural models. 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. These dependencies are acyclic, and have their roots in the exogenous variables. HH only consider structural models with 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 u of the exogenous variables – a so-called *context* – determines a unique world, denoted by s_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 the variables in a world take. 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, their possible values are ranked according to typicality. 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. 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. We will also 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, 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 2. *[HH-extension of actual causation.] Given are an extended structural model $(M, \succeq)$ and context u, such that both C and E hold in s_u. C is an HH-actual cause of E in $(M, \succeq, u)$ iff C is an actual cause of E in (M, u) when we consider only witnesses w such that $w \succeq s_u$.*

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.

4 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.

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 (1) and (2)), while a dependency on only endogenous variables results in a deterministic law (such as (3)).

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 Def. 2 we restrict an extended structural model 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 which have a probability lower than 0.5, and are inconsistent with the assignment in the leaf of b. 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 3. *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 theories with *norms*. Everything that can possibly happen is described by the CP-laws of a theory, hence we choose to introduce *prescriptive* norms as corrections to *descriptive* CP-laws. These corrections 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 (1) with the norm that the professor shouldn't take pens, even though he often does, gives $(Prof : 0.7\{p\}) \leftarrow$, where $1 \gg p$. The *normatively normalized* refinement is then given by the CP-law $(Prof : p) \leftarrow$. 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 $0 < p < 1$, 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 6.4.)

Definition 4. *Given an extended theory T, i.e., a theory also containing norms, we define the* normatively normalized refinement *of T as $T^{NN} := \{r^{NN} | r \in T\}$, where r^{NN} is the normatively normalized refinement of r.*

We can combine both senses of normality, as follows:

Definition 5. *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.[2] In case of our example, $T^{Normal(b)}$ is given by $(Prof : p) \leftarrow$ and laws (5) and (6).

---

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

Remains to be explained 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 Def. 2, 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, and it is at least as normal as the actual world. Similarly, we need to limit the stories allowed by Def. 1 – which are described by T^* – 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 $(T^*)^{Normal(b)}$. Unfortunately we need to treat the unique law that contains C in its head – denoted by $r(C)$ – somewhat special. That's because in Def. 1 the intervention $do(\neg C)$ makes this law deterministic with an empty head. As we want to take into account the (ab)normality of $\neg C$, we work with the normal refinement instead. Hence we denote by T^{**} the theory identical to T^* except that it contains the normal refinement of $r(C)$.

Definition 6. *Given an extended theory T, a story b such that C and E hold in its leaf, and the theory T^* as described in Section 2.2. We define the* normal refinement *of T^* according to b and C as $T^{Normal(b)*} := (T^{**})^{Normal(b)}$.*

Besides the normality and belonging to T^*, recall that the remaining requirement for a story to be a witness, is that $\neg C$ and $\neg E$ hold in it. This leads us to the following formulation of the HH-approach in CP-logic.

Definition 7. *[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 r^C with the deterministic law $Assistant \leftarrow$, leading to $P(\neg NoPens \wedge \neg Assistant) = 0$. On the other hand, for $C = Professor$, the normative normalisation replaces r^C with $(Prof : p) \leftarrow$, leading to $P(\neg NoPens \wedge \neg Prof) = q$. 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 $Prof$.

We now show that Def. 7 is indeed the correct translation of the HH approach from structural models to CP-logic.

Theorem 1. *C is an HH-actual cause of E in an extended model and context $(M, \succeq, u)$ iff C is an HH-CP-logic-actual cause of E in (T, b), where (T, b) is derived from $(M, \succeq, u)$ in the sense described above.*

Proof. See Appendix. □

5 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, Def. 7 tries to circumvent the use of this intervention by simply demanding that $\neg C$ holds. However this solution is not always available, and when it isn't this provides counterintuitive results. 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, also when taking into account the typicality of C. 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 $P_{T^{Normal(b)}}(\neg C)$. As the following theorem shows, our new choice only makes a difference in a limited set of cases.

Theorem 2. *If $r(C)$ is non-deterministic 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. See Appendix. □

If $r(C)$ is deterministic 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 $1 * 0.9 = 0.9$.

6 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.

6.1 Why it is better to use $(T^*)^{NN}$ in the first factor

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 the current definition. 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.

$$(Throw(1,1):0.1) \leftarrow .$$
$$(Throw(2,1):0.1) \leftarrow \neg Throw(1,1).$$
$$(Throw(3,1):0.1) \leftarrow \neg Throw(1,1) \wedge \neg Throw(2,1).$$
$$...$$
$$WinCar \leftarrow Throw(1,1).$$
$$WinCar \leftarrow \neg Throw(2,1) \wedge ... \wedge \neg Throw(100,1).$$

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

$$(Throw(1,1):0.1) \leftarrow .$$
$$WinCar \leftarrow Throw(1,1).$$
$$WinCar \leftarrow \neg Throw(2,1) \wedge ... \wedge \neg Throw(100,1).$$

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 $WinCar$.

On the other hand 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 $WinCar$.

6.2 Why we should look at all witnesses

The current example also illustrates why it makes sense to aggregate the strength of all witnesses: the best witness for $Throw(1,1)$ is the story in which $\neg Throw(1,1) \wedge Throw(2,1) \wedge \neg Throw(3,1) \wedge ... \wedge \neg Throw(100,1)$, with a probability of 0.09, thus making $Throw(1,1)$ a very minor cause of $WinCar$ under the HH-approach. Put informally, the alternative we should be considering is not one particular outcome – the best witness – of the throwing sequence that makes you lose the car, but the set of all such sequences – all witnesses – taken together.

6.3 Why it is better to use T^{NN} in the second factor

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.

6.4 The final definition

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

Definition 8 (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 $P_{(T^*)^{NN}}(\neg E|do(\neg C)) * P_{T^{NN}}(\neg C) > 0$.*

7 Conclusion

Our final definition extends our original definition of actual causation (Def. 1) by incorporating the main points raised by HH: (1) it allows normative considerations and (2) is able to factor in the normality of the cause, which is important when we are considering interventions and explanatory power. In addition, it also improves on the HH account in several ways:

- CP-logic is more expressive than structural models, hence it can be applied to more examples (Vennekens, Denecker, and Bruynooghe 2010).

- 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.

- It is able to deal with all of the examples from the HH-paper equally well as Def. 7.

- It can also handle the previous examples properly, as opposed to Def. 7.

8 Acknowledgements

References

Beckers, S., and Vennekens, J. Towards a general framework for actual causation using cp-logic. *Unpublished*. http://arxiv.org/pdf/1410.7063v3.pdf.

Beckers, S., and Vennekens, J. 2012. Counterfactual dependency and actual causation in cp-logic and structural models: a comparison. In Kersting, K., and Toussaint, M., eds., *Proceedings of the Sixth STAIRS*, volume 241 of *Frontiers in Artificial Intelligence and Applications*, 35–46.

Hall, N., and Paul, L. A. 2003. *Causation and Preemption*. Oxford University Press.

Hall, N. 2004. Two concepts of causation. In *Causation and Counterfactuals*.

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

Halpern, J. Y., and Hitchcock, C. Graded causation and defaults. *British Journal for the Philosophy of Science* forthcoming. http://arxiv.org/pdf/1309.1226v1.pdf.

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

Hitchcock, C., and 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., and Fraser, B. 2008. Causal judgment and moral judgment: Two experiments. In Sinnott-Armstrong, W., ed., *Moral Psychology*. MIT Press.

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

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge University Press.

Shafer, G. 1996. *The Art of Causal Conjecture*. Artificial Management. MIT Press.

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

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