STAIRS 2012
K. Kersting and M. Toussaint (Eds.)

doi:10.3233/978-1-61499-096-3-35

Counterfactual dependency and actual causation in CP-logic and structural models: a comparison

Sander Beckers [1] and Joost Vennekens
Dept. Computer Science, KULeuven, Belgium
Campus De Nayer, Lessius Mechelen, Belgium

Abstract. The solution to the problem of actual causation - i.e. determining what caused an effect in a specific scenario - put forward by Halpern and Pearl recently has received a lot of attention. It forms the basis for many other approaches within the dominant tradition of counterfactual theories of causation. However, their solution runs into a number of difficulties for a certain type of examples exhibiting so-called switching causation and early preemption. We discuss these in the light of the core concept of counterfactual dependency, and offer a comparison with the recent definition of actual causation formulated in CP-logic. We argue both that for this type of examples the CP-logic definition provides better anwers, and that it does more justice to the fundamental intuitions underlying counterfactual dependency.

Keywords. CP-logic, structural models, causation, counterfactual dependence

Introduction

The use of causal models is becoming evermore widespread within science. An important open problem in this context is that of actual causation, which can be formulated thus: 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 thing actually caused another? Lewis [8] was the first to propose a solution, and there has been a lively philosophical debate over the issue since. Recent work by Halpern and Pearl [4] has also garnered interest in this topic in the AI community, due to its relevance for domains such as diagnosis, legal reasoning and natural language understanding. Their account (which we will refer to as HP) constructs a formal definition in the context of structural models [10]. Another approach has been developed by Vennekens [14] within the framework of CP-logic, a Prolog inspired language that in many ways improves upon the structural models context [12,13].

Both frameworks follow the tradition started by Lewis of basing their definition on the notion of counterfactual dependency. A preliminary discussion in [14] argued already that CP-logic's use of an explicitly dynamic semantics and the default/deviant

[1]Research funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT-Vlaanderen)

distinction offered a more detailed picture of actual causation. In this paper we will focus on the different ways in which HP and CP-logic make use of counterfactual dependence, concluding that CP-logic does so in a more elegant and consistent manner.

In Section 1 we clarify the relation between actual causation and counterfactual dependence in terms of possible worlds. Section 2 will briefly recall the semantics of both structural models and CP-logic. This sets the stage for the respective definitions of actual causation in Sections 3 and 4. To assess their merit, we discuss the different ways in which both handle a number of controversial examples regarding switching causation and early preemption in Section 5. The last section contains some conclusions and ideas on future work.

1. Possible worlds and counterfactual dependency

By far the most popular approach towards defining the notion of actual causality is to do so in terms of counterfactual dependency: to find out whether C was a cause of E, we imagine that $\neg C$ holds and see if this would affect the value for E. If it does, then C was a cause for E. It is customary to interpret such counterfactuals in terms of possible worlds, in the following sense: is there a relevant possible world satisfying $\neg C$, in which $\neg E$ holds? Here, we do not use the term *possible world* in the narrow sense of an assignment to all variables of the domain, but we take the broader view that it contains the causal laws of the domain as well. This possible worlds interpretation shall serve as our guideline in discussing questions of actual causation throughout the rest of this paper in an attempt to clarify and settle some of the controversies that surround it.

Every counterfactual approach to actual causation can be characterised completely by formulating it as an answer to the following question: "Given that we are looking at a putative cause C and an effect E, which possible worlds are relevant - i.e. should be taken into consideration - to evaluate if there is a counterfactual dependence?". Within the counterfactual tradition there is agreement that any such answer should satisfy the following two general principles.

First of all, we may only take into account those parts of the actual world that are explicitly given by the causal model and the story that takes place describing the events leading up to E. So one should be careful not to base intuitions regarding putative counterexamples to a definition of actual causation on information which is not explicitly part of the model or the story.

Secondly, our ambition is to investigate the actual relation between C and E by looking at worlds in which C no longer occurs. Of course there are many worlds in which $\neg C$ holds, for example those in which some cause D of C is prevented from happening. But if we were to take into account a world in which $\neg D$ holds, we would in fact be investigating the causal relation between D and E, rather than C and E. Therefore we have good reason to adopt the standard solution of considering possible worlds where C is miraculously prevented from being caused, without there being any causal explanation of why this is so. This is the intuition that lies behind Pearl's important notion of an intervention. Everything else that happens in those possible worlds should then be maximally similar to what actually happened. The previous sentence was deliberately phrased in vague terms, since we simply mean to address the intuition that we should look at worlds in which C no longer occurs that are nearest (i.e., most similar) to ours. Another way of

formulating this fundamental principle is that we should be able to explain any difference between the actual world and some possible world we are looking at solely by invoking the fact that C no longer occurs.

Many of the differences between various counterfactual accounts of actual causation are either due to a violation of the first principle or due to the different ways in which the second principle is fleshed out. By bringing this to the fore, we shall compare the HP account with that of CP-logic.

2. Preliminaries: structural models and CP-logic

Before developing an account of actual causation, we need some kind of causal model. HP use structural models, whereas our account will use the language of CP-logic. In this section, we briefly recall both formalisms. Let us assume that we have a finite set of random variables to describe our domain. For simplicity, we also assume that all these variables are Boolean. Each random variable can be either endogenous (internal to the causal model) or exogenous (on the boundary of the causal model).

A *structural model* M consists of a set of equations of the form $X := \phi$, where X is an endogenous random variable and ϕ is a Boolean formula in the random variables. For each endogenous variable X, M should contain precisely one equation defining X. Moreover, the set of equations should be acyclic in order to ensure that each assignment $\vec{O} = \vec{o}$ of values to the exogenous variables $\vec{O}$ induces a unique assignment $\vec{N} = \vec{n}$ of values to the endogenous variables $\vec{N}$ that satisfies the equations. The semantics of the structural model is then given by all such assignments. Typically, a probability distribution $\mathbf{P}(\vec{O} = \vec{o})$ is defined over the assignments of values to the exogenous variables, which then of course also induces a probability distribution $\mathbf{P}(\vec{O} = \vec{o}, \vec{N} = \vec{n})$ over the whole domain. These assignments represent in a straightforward way the different states that the domain could possibly be in. Moreover, the semantics also ensures in an equally straightforward way that each assignment obeys the causal laws by requiring that each equation is satisfied.

CP-logic, on the other hand, uses semantic objects that contain more information. Instead of only representing the states that the domain could eventually be in, they also represent the ways in which the domain could have reached these states. To represent this evolution of the domain, *probability trees* [11] are used. The root of such a tree represents an initial state of the domain, the edges from a node to its children represent a non-deterministic transition from one state to another, the probabilistic labels of the edges quantify this non-determinism (and therefore the sum of the labels of all outgoing edges from a non-leaf node must always be 1), and the leaves of the tree represent the final states that the domain could reach after starting in the given initial state. Each probability tree $\mathscr{T}$ defines an obvious probability distribution $\pi_{\mathscr{T}}$ over its leaves, namely, the probability $\pi_{\mathscr{T}}(l)$ of a leaf l is the product of the labels of all edges that lead to l.

Each node s of a tree is mapped to an interpretation of the random variables $\mathscr{I}(s)$, which represents the corresponding state of the domain. Given this mapping, the probability distribution over leaves of the tree induces an obvious probability distribution over interpretations (the probability of I is $\sum_{\mathscr{I}(l)=I} \pi_{\mathscr{T}}(l)$) and over Boolean formulas (the probability of ϕ is $\sum_{\mathscr{I}(l)\models\phi} \pi_{\mathscr{T}}(l)$).

CP-logic makes the assumption, common in causal modeling, that each endogenous random variable has a *default* value from which it starts, and a *deviant* value that it will

take on when some causal mechanism acts upon it. For uniformity, it is assumed that the default value of each random variable is **false** and that its deviant value is **true**.

Unlike structural models, CP-logic has two distinct sources of uncertainty. First of all, there is uncertainty about the initial state in which the domain starts out. In this initial state, each of the endogenous variables will still have its default value, but the exogenous variables may have any value. Throughout the evolution of the domain, the value of the endogenous variables may change from default to deviant when a suitable causal mechanism acts upon it, but the value of the exogenous variables is assumed to remain constant. In this way, the values of the exogenous variables at the start of the process basically define the context in which it takes place. The uncertainty over this context is not part of the causal model itself and we will typically not quantify it probabilistically. By contrast, there is also uncertainty that is part of the process itself, namely, every time a node has more than one child, there is uncertainty over which of these children will actually be the "real" next state of the domain. The probability tree of course does quantify this uncertainty, yielding the distribution $\pi_{\mathscr{T}}$. Each branch in a such a tree thus represents a possible evolution of the domain.

Whereas a structural model statically defines a set of possible outcomes by giving a set of equations that they must satisfy, a probability tree dynamically defines a set of possible outcomes (i.e., its leaves) as the final states of a generative process. The goal of CP-logic is to provide a modular syntax for describing such generative processes. The basic "atomic" building blocks that make up such a process are the individual probabilistic transitions from one state to the next. In CP-logic, such a transition is described by a *CP-law* of the following form:

$$(H_1 : \alpha_1) \vee \ldots \vee \cdots \vee (H_n : \alpha_n) \leftarrow \phi,$$

where the H_i are Boolean variables, the α_i real numbers that sum to at most 1 and ϕ a Boolean formula. The meaning of this CP-law is that, in any state s where ϕ holds, the non-deterministic event that is described by the disjunction may happen, i.e., the children of s are nodes $\{s_1, \ldots, s_n\}$, where for each i, the probability of the edge (s, s_i) is α_i and the interpretation $\mathscr{I}(s_i)$ may differ from $\mathscr{I}(s)$ only by having $H_i = \textbf{true}$. In case $\sum \alpha_i < 1$, then s will also have one additional child s_0 such that the probability of (s, s_0) is $1 - \sum \alpha_i$ and the interpretation $\mathscr{I}(s_0)$ is simply identical to $\mathscr{I}(s)$. Intuitively, such a CP-law therefore expresses that ϕ causes at most one of the H_i, and each α_i gives the probability that H_i is the boolean variable that is caused.

CP-logic now represents a causal model by a set of these CP-laws. Such a set is called a *CP-theory* and its semantics is given by a set of probability trees, namely all those that can be constructed using the CP-laws in the way outlined above. The resulting trees are called the *execution models* of the CP-theory.

Note that, while building such an execution model, there may be many CP-laws that could be used in any particular node s (namely, all those whose precondition ϕ is satisfied in $\mathscr{I}(s)$). However, CP-logic does not allow concurrency, so only one of these applicable CP-laws will actually happen in s. In terms of the final outcome of the domain, it does not matter which of them is chosen first, since all laws that are applicable in s will remain applicable in later states. This property is obvious for CP-laws whose precondition does not contain negation, because subsequent interpretations only increase in truth, i.e., the only thing that happens is that more and more variables deviate from their default value.

The semantics of CP-logic takes special measures to ensure that this same property also holds for CP-laws containing negation. This is done by ensuring that a CP-law whose precondition depends on some random variable X still being in its default state can only happen once there no longer exists any way in which X could still be caused to deviate. In other words, it is not enough that X is false in the current state s, but it must actually be the case that X has already become impossible. This is formally defined by means of a fixpoint construction that overestimates everything that is still possible in s. We refer the reader to [12] for the details of this construction. The bottom-line, however, is that it produces a set $\mathscr{U}(s) \supseteq \mathscr{I}(s)$ that contains all Boolean variables for which it is still possible that they could become true in some descendant of s. A law may then only be applied in s, if its body holds in both $\mathscr{I}(s)$ and $\mathscr{U}(s)$, since this means that it is not only true now, but will remain true from now on.

This now ensures that it does not matter which applicable CP-law happens first. All probability trees $\mathscr{T}$ that can be constructed starting from the same interpretation O for the exogenous variables, define precisely the same probability distribution $\pi_{\mathscr{T}}$. For a CP-theory T, we denote this unique distribution by π_C^O.

Of course when we are answering questions regarding actual causation, then the order in which CP-laws were applied does matter. Such information will allow a more specific delineation of the possible worlds we want to consider. As a consequence, a question of actual causation in CP-logic requires one to look not only at a CP-theory T and an assignment $\{\vec{O} = \vec{o}\}$ to the exogenous variables, but also at the branch b representing the order in which the CP-laws were applied.[2] The counterpart in the context of structural models is given by a structural model M and an assignment to the exogenous variables $\{\vec{O} = \vec{o}\}$.[3]

3. Actual causation in CP-logic[4]

As stated earlier, our definition attempts to formalize the intuition behind counterfactual dependency, by explicitly describing the set of possible worlds that should be taken into account given a theory T, an assignment $\{\vec{O} = \vec{o}\}$, a branch b and the fact that we want to know whether C caused E.

We want to make sure that everything which can happen as it actually happened, does so. So for each CP-law r that happens in b with effect A, we construct a deterministic CP-law r^A of the form $(A : 1) \leftarrow \phi$, where ϕ is the precondition of r. Further, in order to prevent the putative cause C from occuring, we remove C from the head of every law containing it. This transforms a law r into $r^{\neg C}$. By applying both types of transformations on all laws in T for which they are applicable, we transform our original theory T into $(T^b)^{\neg C}$ which differs from our initial theory T exactly by the following two properties:

- all the CP-laws that happened in b now deterministically cause the same effect as they had in b;
- C is prevented from occurring.

[2] If we have no information concerning b, the definition of actual causation will look at all possibilities for b. See [14] for more details.

[3] The fact that there is no counterpart of a branch b in the structural models approach is due to its static framework. This important difference is discussed as well in [14].

[4] The definition we present here was introduced in [14]

However, the possible worlds that can be generated with this theory still contain a large degree of freedom concerning those CP-laws that weren't applied in the actual world by the time E occurred. We can separate these laws into two categories:

1. laws that were still possible when E occurred;
2. laws that had become impossible at the point when E occurred (i.e. their bodies were false in $\mathscr{U}(s)$).

CP-laws belonging to the first category describe causal mechanisms that did not interfere with the actual course of events leading up to E, although they could have. If such a law happens in a possible world in which $\neg C$ holds, then this is a difference with the actual world that cannot be explained solely by invoking the fact that C fails to hold, since it was a viable option in the actual world where C *does* hold. Therefore allowing this type of laws would go against our second principle, so we deem them as irrelevant with regards to counterfactual dependency and do not take them into consideration.

The situation is different for the second category of CP-laws. Assume r is a law that would become possible just in case $\neg C$ holds. Then r describes a kind of causal mechanism that might be set in motion precisely because C does not occur, and it is clear that any possible world we are taking into consideration should contain it. If, however, r will remain impossible even when $\neg C$ occurs, then it shall never be applied and it is therefore harmless to keep it in our theory.

This leads to the following definition.[5]

Definition 1 (Actual causation) *Given a theory T, a branch b such that both C and E hold in its leave, and an initial interpretation O. To find out if C is an actual cause of E, you look at b, find the place where E first appeared, discard all events that had not yet happened then but still were possible, transform the remaining theory T' into $T'' = (T'^{b})^{\neg C}$ and check whether $\pi^{O'}_{T''}(\neg E) > 0$.*[6]

In other words, C is an actual cause of E if there is a possible world in which $\neg E$ holds and:

- C is prevented from occurring;
- whenever a CP-law is applied that was actually applied before E occurred, it has the same effect as it actually had;
- whenever a CP-law is applied that wasn't actually applied by the time E occurred, it had actually become impossible at the point when E occurred (i.e. their bodies were false in $\mathscr{U}(s)$).

So far, we haven't considered causation by omission ("did the doctor's failure to treat the patient cause his death?"), and negative effects ("did the doctor's treatment prevent the patient's death?"). Extending the framework to also address such questions is easy enough:

- To extend our definition of actual causation to allow also literals $\neg E$ to act as effects, we need to specify when such a $\neg E$ "happens" for the first time, such that we may discard all later events when making counterfactual judgments to deter-

[5]For a formal version, see [14].

[6]To cover the case where C is exogenous, O' is $O \setminus \{C\}$.

mine what caused $\neg E$. The obvious cut-off point is when E no longer belongs to the overestimate $\mathscr{U}(s)$.
- To also allow literals $\neg C$ to act as causes, we need to define precisely how we will check the counterfactual dependency in this case. To assume that $\neg C$ was not the case, we need to assume that C has somehow occurred, which we can do formally by just adding a new CP-law "$C \leftarrow$" that always causes C.

4. Actual causation in HP

In order for this paper to be self-contained we state the formal version of the HP account in this section, but due to space restrictions we do not go into it here. The interested reader may look at [4] for a thorough explanation of it. In this definition, M is a structural model with endogenous RVs $\mathscr{V}$, $\vec{u}$ an assignment of values to the exogenous RVs, $\vec{X}$ a tuple of endogenous RVs, and ϕ a Boolean formula in the RVs. The notation $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}]\phi$ means that ϕ holds in $(M,\vec{u})$ after the intervention of assigning $\vec{x}$ to $\vec{X}$ is performed, i.e., each $X_i \in \vec{X}$ has its defining equation removed from M and replaced by $X_i := x_i$.

Definition 2 (HP account of actual causation) *$\vec{X} = \vec{x}$ is an actual cause of ϕ in $(M,\vec{u})$ if the following three conditions hold.*

AC1. $(M,\vec{u}) \models (\vec{X} = \vec{x}) \wedge \phi$. *(That is, both $\vec{X} = \vec{x}$ and ϕ are true in the actual world.)*

AC2. There exists a partition $(\vec{Z},\vec{W})$ of $\mathscr{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}',\vec{w}')$ of the variables in $(\vec{X},\vec{W})$ such that if $(M,\vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\phi$. *In words, changing $(\vec{X},\vec{W})$ from $(\vec{x},\vec{w})$ to $(\vec{x}',\vec{w}')$ changes ϕ from true to false.*

(b) $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\phi$ *for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$. In words, setting any subset of variables in $\vec{W}$ to their values in $\vec{w}'$ should have no effect on ϕ, as long as $\vec{X}$ is kept at its current value $\vec{x}$, even if all the variables in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.*

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2. Minimality ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing ϕ in AC2(a) are considered part of a cause.

The heart of this definition lies in AC2, which is meant to capture counterfactual dependence. Here, the interventions $\vec{W} \leftarrow \vec{w}'$ represent changes to the actual setting beyond an intervention on the putative cause $\vec{X} = \vec{x}$. The idea is, roughly, that $\vec{X} = \vec{x}'$ is not sufficient to guarantee the truth of $\neg\phi$ by itself, and other interventions may be required, but $\vec{X} = \vec{x}$ is sufficient to make sure ϕ holds in the current context.

There are three fundamental differences between our account of actual causation and that of HP:

- CP-logic makes use of a default/deviant distinction;
- the dynamic nature of causality is an explicit part of CP-logic and not of the structural models framework;

- HP takes all causal mechanisms into consideration in possible worlds, but makes use of contingencies to counteract irrelevant mechanisms, whereas CP-logic disposes of irrelevant CP-laws to do so.

The first two differences have been discussed in [14] and are not specifically related to the topic of possible worlds. We will focus on the third, which is fundamental to the examples we now turn to.

5. Switching causation and early preemption

Consider the following example of early preemption from Hitchcock [7].

> **Assassin.** Assassin poisons Victims coffee, Victim drinks it and dies. If Assassin hadn't poisened the coffee, Backup would have, and Victim would have died anyway. Victim would not have died if there had been no poison in the coffee.

The obvious formalization in CP-logic is:

$$Assassin \leftarrow . \tag{1}$$

$$Backup \leftarrow \neg Assassin. \tag{2}$$

$$Dies \leftarrow Assassin \vee Backup. \tag{3}$$

Both HP and Hitchcock designate *Assassin* as the actual cause of *Dies*, whereas CP-logic does not. The reasoning for CP-logic is as follows: all three laws are relevant, as well as deterministic, so we simply need to check for a counterfactual dependence of *Dies* on *Assassin* in the theory as it is. Since there isn't any, *Assassin* is judged not to be a cause of *Dies*. The HP definition answers otherwise because it considers the contingency that *Backup* is false no matter whether *Assassin* holds or not, in which case there is a counterfactual dependence.

HP and Hitchcock base their answer on its intuitiveness: we naturally tend to think of *Assassin* as the cause of *Dies*, even though there is a backup causal mechanism that will lead to the same effect in case *Assassin* fails to do his job. Before giving our response to this argument, it will be helpful to consider the next example from Hall [3], that shares the same formal properties as the previous one.

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

The causal model behind this story can be represented in CP-logic as follows.

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

$$LeftTrack \leftarrow Switch. \tag{5}$$

$$RightTrack \leftarrow \neg Switch. \tag{6}$$

$$Destination \leftarrow LeftTrack \vee RightTrack. \tag{7}$$

Since this example is formally identical to the previous one, HP designates the flipping of the switch as a cause of the train arriving at its destination. However, intuitively we feel that directing the train from one track to another that serves exactly the same purpose is not a cause of its arrival. So in this case CP-logic has intuition on its side.

HP try to defend their judgment as follows.

> Is flipping the switch a legitimate cause of the trains arrival? Not in ideal situations, where all mechanisms work as specified. But this is not what causality (and causal modeling) are all about. Causal models earn their value in abnormal circumstances, created by structural contingencies, such as the possibility of a malfunctioning track. It is this possibility that should enter our mind whenever we decide to designate each track as a separate mechanism (i.e., equation) in the model and, keeping this contingency in mind, it should not be too odd to name the switch position a cause of the train arrival (or non-arrival).

If we reformulate this argument in terms of possible worlds, it states that we should not only look at possible worlds where (5) fails to work, but also take into consideration possible worlds in which (6) fails as well. Clearly the latter are strictly further away from the actual world than the former, and taking them into consideration violates the second principle we stated earlier.

A much simpler solution is possible within CP-logic. HP base their argument on the possibility that the causal mechanism for *RightTrack* breaks down. In CP-logic this would come down to making (6) non-deterministic, modelling the fact that railroad tracks can break down every now and then. In that case, *Switch* would be counted as a cause of *Destination*, and the two definitions would agree. Unfortunately there is no way of modelling such indeterminism directly in the structural models account, and therefore HP cannot make such a nuanced distinction.

Comparing the *Switch* example with the *Assassin* one, we offer two observations in defense of our account.

First, it is important to note that in the given model the backup causal mechanism exhibited by (2) is deterministic: we have absolute certainty that *Backup* will fulfill his job if *Assassin* fails. As HP's argument regarding the *Switch* confirms, our basic intuition in these type of examples often implicitly assumes a non-deterministic mechanism, which is in violation of our first principle. A more realistic model would add at least a fraction of indeterminism to (2), in which case CP-logic would give the same answer as HP and Hitchcock do. The fact that this is more outspoken in the *Assassin* example, could be explained by noting that in general we are more confident in the workings of railroad tracks than we are in the actions of people.

Second, although appeals to intuition with regards to concrete examples of actual causation should play an important role in assessing any formal definition of the concept, their importance should not be overestimated. Not only is there the evident problem that people do not share the same intuitions on every example, but more importantly, even a single person's intuitions regarding one example often conflict with those regarding another that has the same formal properties, as the previous two examples show. All one can do is make the most useful and consistent choice between conflicting intuitions.

5.1. Intuitions and causality

It is worthwhile elaborating on the previous point, since there is an abundance of appeals to the intuitiveness of causal statements in the literature on actual causality. We will do so by a drawing a parallel to the problem of *moral luck*, and by means of two examples that are formally identical to the ones discussed above.

It is a well-known issue in moral philosophy that people attach much more importance to what actually happened in the case of ethical situations, than they do in life in general. (See [9,15] for details.) That is to say, given that two identical courses of action produce different outcomes, moral judgments concerning each can differ considerably. The next two examples illustrate this point.

> **Train.** Both passengers A and B take the train an hour ahead of schedule to catch the same flight. However, passenger A's train runs into a delay of two hours, causing him to miss his flight. Passenger B arrives on time.

It seems rather uncontroversial to say that both passengers A and B were equally well prepared, and acted equally rational. Passenger A simply had bad luck, but we will not think less of him because of it. Things are different in the next example.

> **Moral luck.** Both drivers A and B are drunk. Both of them run a red light. In both cases, a child is crossing the street. The child in the case of driver A is looking towards the car and jumps away just in time. The child in the case of driver B, however, doesn't see the car coming and gets killed in the accident.

Our moral assessment of driver B's deeds will be far more severe than those of driver A, and the corresponding legal consequences would confirm this. But when we look at their actions purely from an instrumental point of view, their actions were identical and thus they should be judged in the same manner. In the Train example we did resort to such purely instrumental intuitions, since there was no ethical content.

We will now present two examples to illustrate that a similar divergence between an instrumental and an evaluative perspective is at work in our examples on causality.

> **Euthanasia.** A doctor injects a lethal dose of morfine into a terminally ill patient, who would have died with certainty within the next three months.

Of course the injection is a cause of the patient dying at this time point, rather than just somewhere within the next three months. But should we count it as a cause of the patient's dying within the next three months? As the intense debates surrounding euthanasia show, intuitions on this subject are divided.[7] Moreover, since euthanasia is legal in the Benelux and it is illegal in the USA, there might also be cultural differences underlying the difference in intuitions.

The following formally identical example would not lead to such fierce debates.

> **Barcelona.** With only one game left this season, Barcelona has 4 points more than the runner-up. They then win their last match, gaining 3 points, and become champion. A draw would have given 1 point, and a loss would have given none.

[7] This debate obviously isn't limited to the question of causality, and one may very well share intuitions on causality whilst disagreeing on euthanasia. But many arguments pro and contra will contain causal statements.

Is Barcelona's winning their last match a cause of them becoming champion? According to HP it is, but most people would answer this in the negative, reasoning as follows: "It was already certain that they would have become champion no matter whether they won or not, so the victory is irrelevant". The difference between both examples lies completely in their content, not in their form. The disagreement concerning the first is due to its ethical nature, whereas the consensus regarding the second is due to its deterministic and instrumental character.

If we generalize this message we can conclude that, at least for this kind of examples, a purely instrumental perspective will give consistent answers, whereas any perspective that wishes to take into account evaluative judgments as well is destined to be inconsistent. CP-logic offers an instrumental definition that can handle these examples consistently, whereas HP and Hitchcock can not.

The point we are trying to make here is not that there is something wrong with our moral judgments, and that we should abandon them out of fear for inconsistency. That would be a matter of putting the cart before the horse, logical requirements should answer to moral requirements and not vice versa. Rather, our point is that within the context of a formal definition of actual causality, which is in the first place meant to aid us in improving our factual knowledge of the world, a consistent instrumental perspective is to be preferred over an inconsistent perspective that aims to capture evaluative intuitions as well.

5.2. Counterexample to Hitchcock

Hitchcock [6] relies on contingencies in a similar manner as HP does, and he does so in order to deal with examples of early preemption such as the *Assassin* case.[8] He admits, however, that this type of definition runs into trouble regarding other examples, one of which he gives himself under the heading "Counterexample to Hitchcock" [7].

> Assistant Bodyguard puts a harmless antidote in Victim's coffee. Buddy then poisons the coffee, using a type of poison that is normally lethal, but which is countered by the antidote. Buddy would not have poisoned the coffee if Assistant had not administered the antidote first. Victim drinks the coffee and survives.

This translates into the following CP-logic theory.

$$(Assistant : 1) \leftarrow . \tag{8}$$

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

$$Dies \leftarrow Buddy \wedge \neg Assistant. \tag{10}$$

According to Hitchcock's definition *Assistant* is a cause for Victim's survival, which he contrasts with the observation that "Many people, but by no means all, have the intuition that Assistant's adding the antidote to the coffee is not a cause of Victim's survival", illustrating our previous point about intuitions. The reader can verify that HP gives the same answer as Hitchcock, whereas CP-logic does not.

[8] Although his approach makes use of the notion of an "active path", rather than structural contingencies.

6. Conclusion and future work

The dominant tradition regarding actual causality aspires to come up with a formal definition of the concept in terms of counterfactual dependency. The latter notion is itself usually explained by means of a possible worlds semantics. We have therefore tried to evaluate two important accounts of actual causation by reviewing a class of difficult examples in light of this semantics, by examining how each definition deals with them. Looking into the intuitions behind the examples, an analogy arose with the ethical problem of moral luck, that led to the distinction between an instrumental and an evaluative perspective.

As a result we conclude that the HP account fails to be consistent with intuition for examples involving determinism and/or a purely instrumental context, and it takes into account possible worlds that violate two fundamental principles regarding counterfactual dependency. CP-logic, on the other hand, proves to be consistent with an instrumental view point, and offers an account of counterfactual dependency that stays true to both fundamental principles.

In the future we intend to apply our possible worlds evaluation on a broader set of controversial examples, most notably those involving overdetermination. We shall make some small improvements to the CP-logic definition of actual causation in order to deal with these in a satisfactory manner, and we shall again contrast this with the HP account.

References

[1] GLYMOUR, C., DANKS, D., GLYMOUR, B., EBERHARDT, F., RAMSEY, J., SCHEINES, R., SPIRTES, P., TENG, C. M., AND ZHANG, J. 2010. Actual causation: a stone soup essay. *Synthese 2*, 169–192.

[2] HALL, N. 2004. Two concepts of causation. In *Causation and Counterfactuals*.

[3] HALL, N. 2007. Structural equations and causation. *Philosophical Studies 132,* 1, 109–136.

[4] 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.

[5] HIDDLESTON, E. 2005. Causal powers. *British journal for the philosophy of science 56*, 27–59.

[6] HITCHCOCK, C. 2001. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy 98*, 273–299.

[7] HITCHCOCK, C. 2007. Prevention, preemption, and the principle of sufficient reason. *Philosophical review 116,* 4, 495–532.

[8] LEWIS, D. 1973. Causation. *J. of Philosophy 70*, 113–126.

[9] NAGEL, T. 1979. Moral luck. In *Mortal questions*.

[10] PEARL, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[11] SHAFER, G. 1996. *The art of causal conjecture*. MIT Press.

[12] 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.

[13] VENNEKENS, J., DENECKER, M., AND BRUYNOOGHE, M. 2010. Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In *JELIA*. 313–325.

[14] VENNEKENS, J. 2011. Actual causation in CP-logic. *Theory and Practice of Logic Programming 11*, 647–662.

[15] WILLIAMS, B. 1981. *Moral luck.* In *Moral luck*.