Causal reasoning in a logic with possible causal process semantics

Marc Denecker† and **Bart Bogaerts**† and **Joost Vennekens**‡
firstname.lastname@cs.kuleuven.be
† KU Leuven, Department of Computer Science, Celestijnenlaan 200A, 3001 Heverlee, Belgium
‡ KU Leuven, Department of Computer Science, Campus De Nayer, 2860 Sint-Katelijne-Waver, Belgium

Abstract

We point to several kinds of knowledge that play an important role in controversial examples of actual causation. One is knowledge about the causal mechanisms in the domain and the causal processes that result from them. Another is knowledge of what conditions trigger such mechanisms and under what conditions it can be preempted.

We argue that to solve questions of actual causation, such knowledge needs to be made explicit. To this end, we develop a new language in the family of CP-logic, in which causal mechanisms and causal processes are formal objects. We then build a regularity-theoretic framework for actual causation in which various notions of actual causation are defined. Contrary to counterfactual definitions, actual causes are defined directly in terms of the (formal) causal process that causes the possible world.

Introduction

Since the days of Hume (1739), causal reasoning has been an active research domain in philosophy and (later) knowledge representation. With the groundbreaking work of Lewis (1973) and Pearl (2000), the structural equations and counterfactual reasoning approach became mainstream (Halpern and Pearl 2005; Halpern 2016a; Fenton-Glynn 2015; Gerstenberg et al. 2015). But the debate remains intense (Glymour et al. 2010). The counterfactual approach is contested by some (Hall 2004; Baumgartner 2013; Bochman 2018). In many scenarios, there is no agreement of what are the actual causes, and all definitions of actual causation have scenarios where they have been criticized. It shows that the informal notion of actual causation is vague and overloaded with many intuitions; also that many sorts of knowledge influence our judgment of actual causation. Science is not ready yet with unraveling all this.

Of the most striking examples are those where for the same formal causal model, different informal interpretations can be proposed that lead to different actual causes. Such examples are interesting since they are clear cases that *some* relevant knowledge is missing in the causal model. A powerful illustration is given by Halpern (2016b), who discusses 6 causal examples from the literature in which authors had shown (often convincingly) that the actual causation definition of Halpern and Pearl (2005), henceforth called HP, failed to predict the actual causes. He responds by proposing for each example an alternative informal interpretation leading to the *same* structural equation model but to intuitively *different* actual causes which, moreover, are those derived by HP! Halpern concludes that, as far as actual causation goes, the structural equation models are ambiguous. As for what knowledge is missing, he claims:

> "what turns out to arguably be the best way to do the disambiguation is to add [...] extra variables, which [...] capture the **mechanism of causality**".

That is, Halpern argues that it is necessary to make knowledge of causal mechanisms explicit.

That such information is relevant for causal reasoning is not surprising. Almost every causal scenario in the literature comes with an informal specification of causal mechanisms and a *story* specifying which mechanism are active and how they are rigged together in a causal process. As observed before (Glymour et al. 2010; Vennekens 2011), most of this information is abstracted away in structural equations. We illustrate to what problems this may lead with a simple example. Consider two scenarios involving two deadly poisons, arsenic and strychnine. In the first scenario, intake of any of these poisons triggers a specific deadly biochemical process. The structural equation of this scenario is:

$$Dead := Arsenic\_intake \vee Strychnine\_intake$$

If both poisons are taken, this is an instance of overdetermination; HP derives the intuitively correct answer that both poisons are actual causes of death. The second scenario is similar, except that arsenic, in addition to poisoning the victim, also *preempts* the chemical process by which strychnine poisons the victim. Now, the structural equation remains the same (i.e., the victim dies as soon as at least one poison is ingested) and so do the *possible worlds*! However, the judgments of actual causation differ: when both poisons are ingested, only arsenic is a cause of death, since the effects of the strychnine are preempted. The conclusion is that the structural equation correctly predicts the possible worlds but does not contain enough information to explain the actual causes. The missing knowledge is what the separate causal mechanisms are and when they are active.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The following scenario, simplified from **Assassin (Hitchcock 2007)**, illustrates another relevant sort of knowledge that is not expressed in structural equation models. *An assassin may kill a victim by administering deadly poison. A bodyguard may rescue the victim by administering an antidote.* Consider the following structural equation.

$$Dead := Poison\_intake \wedge No\_antidote\_intake$$

While it correctly characterizes the possible worlds of this domain, there is again a problem on the level of actual causes. When only poison is ingested, there is a strong intuition that it is the ingestion of poison that is the actual cause of death, not the absence of antidote. After all, it is the poison that activates the poisoning mechanism, not the absence of antidote. Yet, by the symmetry of the formal model, HP nor any other mathematical method can discover this from the above structural equation. The asymmetry here is that poison *triggers* the causal mechanism, while antidote *preempts* it, i.e., absence of antidote is only a condition to not *preempt* the mechanism. As we will argue below, this distinction plays a role in many controversial causal examples. Such information is missing and should be added to the causal model.

Halperns solution to the above sort of problem is to *reify* the causal mechanism by an auxiliary variable representing whether it *fires*, and incorporating these variables at suitable places in the theory. While his solutions work, there is room here for a complementary approach in which causal mechanisms and their triggers and preemptors are explicit in the causal model.

We proceed as follows. We first define a formal logic to express this knowledge. The logic gives a syntactical and semantical account of causal mechanisms, the causal processes to which they lead, and the possible worlds that these processes produce. Then we define different notions of actual causation in terms of the causal process that causes the possible worlds. We exploit the fact that a causal process gives a precise *explanation* of the possible world that it causes, from which various notions of causation can be "read off". This results in a framework of regularity-theoretic definitions of actual causation. Then, . . . "calculemus!": we evaluate the approach in several examples.

The causal logic: syntax and informal semantics

The logic below is propositional. To represent a causal domain, a *vocabulary* Σ of propositional symbols is to be designed, each expressing a proposition in the domain. As usual, literals over Σ are formulas of the form P or $\neg P$, with $P \in \Sigma$; slightly abusing notation, we use $\neg L$ to denote P if $L = \neg P$ and to denote $\neg P$ if $L = P$.

A causal theory consists of causal laws: abstract representations of causal mechanisms. Each mechanism has triggering conditions, which set the mechanism in operation, no-preemption conditions, which if false, preempt the mechanism, and an effect. This leads to the following definition.

Definition 1. *A* causal law *is a statement of the form:*

$$L \leftarrow T \,||\, C$$

where

- $\leftarrow$ *is the causal operator (not material implication),*
- L *is a literal,*
- T *is a sequence of literals called triggering conditions,*
- C *is a sequence of literals called no-preemption conditions.*

Elements of $T \cup C$ *are called* conditions *of the causal law.*

A causal theory Δ *is a set of causal laws subject to two constraints:*

- Δ *is acyclic, i.e., there exists a strict well-founded order on symbols such that for each causal law, the symbol in the head is strictly larger than each symbol in the body.*
- Δ *does not contain laws* $P \leftarrow \ldots$ *and* $\neg P \leftarrow \ldots$ *for the same symbol* P*.*

The causal logic so far serves to describe causal mechanisms. We extend it to make it suitable to express observations of the world.

Definition 2. *An* extended causal theory *is a set of causal laws and propositional formulas over* Σ*.*

Example 1. Arsenic and Strychnine *The two causal scenarios mentioned in the introduction are represented as follows:*

$$\left\{ \begin{array}{l} Dead \leftarrow Arsenic\_intake \,|| \\ Dead \leftarrow Strychnine\_intake \,|| \end{array} \right\}$$

(both rules have the empty sequence of no-preemption conditions), and

$$\left\{ \begin{array}{l} Dead \leftarrow Arsenic\_intake \,|| \\ Dead \leftarrow Strychnine\_intake \,||\, \neg Arsenic\_intake \end{array} \right\}$$

As for the last rule, strychnine poisoning is triggered by strychnine but preempted by the presence of arsenic.

An extended causal theory is obtained by adding the formula $Dead$ *expressing the observation that the victim died, or the formula* $Arsenic\_intake \vee Strychnine\_intake$ *to express that at least one poison was ingested.*

As usual, we distinguish between *endogenous symbols* (those in the head of laws) and *exogenous* symbols (the rest). A causal theory expresses *all* causal mechanisms affecting an endogenous symbol and it ignores all those of exogenous symbols. The language is not designed for the epistemic state where only part of the causal mechanisms of an endogenous symbol are known.

In a causal law $L \leftarrow T \,||\, C$, T and C represent conjunctions of literals. A causal mechanism that is triggered by a disjunction of events cannot be expressed by a single causal law, but must expressed using multiple causal laws. This is a limitation from a KR point of view, but it was done to simplify the definition of causal process.

A rule $L \leftarrow ||$ with empty sequences T, C expresses an unconditional causal mechanism causing L.

Definition 3. *A* world W *informally represents a state of affairs; formally, it is a complete and consistent set of literals, i.e., a set of literals such that for each symbol* $P \in \Sigma$*, either* $P \in W$ *or* $\neg P \in W$*, but not both. The* exogenous state *of* W *is the set of its exogenous literals, denoted* $Exo(W)$*.*

Definition 4. *A condition (i.e., a triggering condition or no-preemption condition) K of a causal law r* blocks *r (or is a blocking literal of r) in world W if $\neg K \in W$. We say that r is blocked by K in W. A causal law $L \leftarrow A \,||\, B$ is* active *in world W if $A \subseteq W$, that is, if all its triggering conditions hold in W; otherwise it is inactive. A causal law is* applicable *in W if $A \cup B \subseteq W$. A causal law* is causally preempted *in W if it is active but is blocked by one of its no-preemption conditions in W. A causal law $r = L \leftarrow \ldots$ is* satisfied *in W if it is blocked by some condition, or if it is applicable and its head L holds in W.*

Distinguishing causal mechanisms In many causal domains, properties may be affected by multiple causal mechanisms. E.g., a window may be shattered by one of multiple throws and many more events. A forest fire can be ignited by matches of campers or by lightnings, etc. The modularity principle of knowledge representation suggests to represent such separate objects by separate expressions and so, it is a good thing of the logic that it supports this. Furthermore, this distinction of mechanisms is sometimes needed for determining actual causes, as argued in the introduction.

Even now, before having defined a formal semantics, it is intuitively clear how to transform causal theories to structural equations, namely by *predicate completion* (Clark 1978). E.g., the completion of the first causal theory of **Arsenic and Strychnine** is the propositional logic representation of the structural equation:

$$Dead := Arsenic\_intake \vee Strychnine\_intake$$

The completion of the second theory is syntactically different but logically equivalent.

$$\begin{aligned} Dead := \; & Arsenic\_intake \vee \\ & (Strychnine\_intake \wedge \neg Arsenic\_intake) \end{aligned}$$

The transformation abstracts away the causal mechanisms and the distinction between triggering conditions and no-preemption conditions. We return to this later.

Definition 5. *An endogenous literal L is* deviant *in Δ if it is the head of a causal law. Otherwise L is* default*. A symbol P is in its deviant state in world W if its deviant literal holds in W; otherwise it is in its default state.*

The distinction between deviant and default literal is not made in structural equations but is found in several other formalisms (Hall 2007). The idea is that a symbol is in its default state unless some causal mechanism brings it in its deviant state. A deviant literal L that holds in the world is *caused* by at least one causal mechanism. A default literal L that holds in the world has a *reason*, namely, that every causal mechanism that can cause the deviant literal $\neg L$ is blocked. Either way, the logic implements Leibniz's principle of sufficient reason –that nothing happens in the world without a cause– for deviant as well as for default endogenous literals (but not for exogenous literals).

In many causal domains, causal mechanisms exist that make a property true and others that make it false. E.g., flipping a switch causes the light to be on if the light was not on (off) and vice versa. In our logic, it is not possible to express both mechanisms in the same theory. It is a limitation that, in our opinion, is inherent to the non-temporal nature of the logic. We argue that such combinations of mechanisms are useful mainly in a setting where the truth of propositions fluctuates in time. Like most languages in this area, the logic proposed here is not equipped for modeling such situations.

Triggering conditions versus no-preemption conditions The distinction between triggering conditions and no-preemption conditions of causal mechanisms is a new feature of our logic. Often, a natural distinction can be made between the conditions that set the mechanism in operation and conditions that are necessary for the mechanism to succeed. E.g., to obtain a forest fire, at least three conditions are needed: a forest, a spark igniting a hotbed and absence of extinction operations. It is only the spark (in the form of a lightning or an unsafe camp fire) that triggers the causal mechanism (triggering condition). We argue that this explains the strong intuition shared by many that it is the spark that is the actual cause of the fire, and not the existence of the forest or the absence of fire extinction. We find the same distinction in many examples. The combination of presence of poison in a coffee and drinking it activates a biochemical process that kills the victim (triggering condition) unless blocked by an antidote (no-preemption condition). Suzy's throw at a window activates a process that shatters the window (triggering condition) if her shoulder is not soar (which prevents throwing hard enough), if the stone is not intercepted, if the window is still intact (no-preemption conditions). In each case, we perceive a distinction between triggering conditions and no-preemption conditions. This appears to affect our judgment of the actual causes: in at least one view of actual causation, the triggering conditions are actual causes of the effect, while the no-preemption conditions are not. Hence, to derive this notion of actual causation, the nature of the conditions must be clear from the causal theory.

Example 2. *(***Drinking poisoned coffee***, see (Hitchcock 2007)) Drinking poisoned coffee causes death unless an antidote is administered. There are three conditions here: presence of poison in the coffee ($Poison$), drinking the coffee ($Drink$), and absence of antidote ($\neg Antidote$). It is clear that $\neg Antidote$ is a no-preemption condition but what about $Poison$? The poisoning process is physically triggered by the event of drinking; the poisoning of coffee could have taken place long before. How could $Poison$ be a triggering condition then? On the other hand, it is still the intake of poison that triggers the poisoning process. This condition, in the chosen vocabulary, is best represented by the conjunction of $Drink$ and $Poison$. So, we argue for the following representation:*

$$\neg Alive \leftarrow Drink, Poison \,||\, \neg Antidote$$

Under this representation, it will be derived that $Drink$ and $Poison$ are actual causes of $\neg Alive$, but $\neg Antidote$ is not.

The above example raises a subtle concern. The scientific goal of actual causation research is to find methods to solve

actual causation problems by *deriving* actual causes and preemptions. But in Example 2, one might get the impression that we are directly *encoding* the desired solution of the actual causation problem in the causal model. However, this is not what we do. In the above example, as in many other causal domains, there is a strong and clear intuition of what triggers causal mechanisms and what may preempt them. The distinction is independent of the specific actual causation problem and is relevant in many different actual causation problems. E.g., in **Drinking poisoned coffee** (Ex.2), the fact that $Drink$ and $Poison$ are triggering conditions and $\neg Antidote$ a no-preemption condition is relevant not only for determining the actual causes of death when all conditions hold; it is relevant as well in the seven other exogenous contexts. E.g., in a context where the victim drinks unpoisoned coffee and ingests an antidote, the common intuition is his survival is caused by the absence of poison, not the presence of antidote. In more complex causal models with many variables and causal laws, the information about triggering conditions or no-preemption conditions of one mechanism may influence actual causation of many variables in an exponential number of exogenous states. This can be seen in double preemption examples.

Formal semantics: causal processes and possible worlds

The formal semantics specifies for each causal theory Δ its causal processes and the world that each process leads to. Causal processes can be formalized in multiple ways. Vennekens, Denecker, and Bruynooghe (2009) formalize it as a sequence of states in which at every state one causal laws is applied until all causal laws are satisfied. This representation is precise and gives an account of, e.g., the "stories" in many causal examples. However, for explaining actual causes, it is a bit too detailed. E.g., it fixes the order of application of causal mechanisms which is largely irrelevant for determining actual causes. So, we opt to formalize a process as an acyclic dependency graph of the firing causal laws.

Definition 6. *A possible causal process for Δ is a directed labeled graph $\mathcal{P}$ on a world denoted $World(\mathcal{P})$. Each arc from literal K to literal L is labeled with a rule r and is denoted $L \overset{r}{\leftarrow} K$ (or $K \overset{r}{\rightarrow} L$). The graph satisfies the following conditions:*

- *For each deviant endogenous literal $L \in World(\mathcal{P})$, there exists a nonempty set F_L of applicable rules with head L, called the* firing rules *of L, such that for each condition K of each rule $r \in F_L$, there is an arc $L \overset{r}{\leftarrow} K$. There are no other arcs to L. We call them* active arcs *and distinguish between* trigger arcs *and* no-preemption arcs *depending on the type of the condition K in r.*
- *For each default endogenous literal $L \in World(\mathcal{P})$, for each rule $r = \neg L \leftarrow \ldots$, the set B_r of blocking conditions of r in $World(W)$ is non-empty and there is an arc $\neg L \overset{r}{\leftarrow} \neg K \in \mathcal{P}$ for each $K \in B_r$. There are no other arcs to L. We call such arcs* blocking arcs *and we distinguish between* no-trigger arcs *and* preemption arcs.

The leafs of a causal process are exactly the true exogenous literals of the world; the non-leafs are the true endogenous literals.

We observe that causal processes can have multiple "simultaneous" causal mechanisms causing the same deviant literal L. That is, L's firing set F_L may contain more than one rule. This is needed to model *overdetermination*.

The active arcs in a causal process reflect the conditions that helped to trigger a causal mechanism causing a deviant literal. The blocking arcs for a rule r reflect all conditions that prevented a causal mechanism to be applicable. A false deviant literal L has at least one blocking arc for every causal mechanism that could cause L.

The difference between triggering conditions or no-preemption conditions in a causal theory barely affects the causal processes and is only visible in the classification of the arcs (trigger, no-preemption, no-trigger or preemption arcs). These labels do not play a role in determining the possible world that the process causes but they will play a key role in the definition(s) of actual causation.

Definition 7. *A causal process $\mathcal{P}$* realizes *world W if $W = World(\mathcal{P})$. We call W a possible world of Δ if it is realized by some causal process for Δ.*

The above notions generalize naturally to extended causal theories. A process $\mathcal{P}$ is a possible causal process for an extended causal theory Δ if it is a causal process for the set of causal laws in Δ and $World(\mathcal{P})$ satisfies the propositional formulas of Δ.

A possible world semantics induces the notions of satisfiability and entailment.

Definition 8. *We say that an (extended) causal theory Δ is* satisfiable *if it has at least one possible world. It* logically entails *a propositional formula φ if φ is true in every possible world of Δ.*

Definition 9. *We say that a causal mechanism r is* chronologically preempted *in $\mathcal{P}$ if it is applicable but does not belong to the fire set of its effect.*

Example 3. *(***Drinking poisoned coffee, cont.***) Each of the eight exogenous states of this causal theory determines a unique process. E.g., the context $\{Drink, Poison, \neg Antidote\}$ is the only context in which the victim dies. The causal law is active and fires and $\neg Alive$ has incoming trigger arcs from $Poison$ and $Drink$ and a no-preemption arc from $\neg Antidote$. In context $\{Drink, Poison, Antidote\}$, the law is active but preempted; $Alive$ has an incoming preemption arc from $Antidote$. In $\{\neg Drink, \neg Poison, Antidote\}$, the rule is inactive and $Alive$ has no-trigger arcs from $\neg Drink, \neg Poison$ and a preemption arc from $Antidote$. The latter context corresponds to* **Bogus Prevention** *(Hiddleston 2005; Hall 2007).*

In many preemption examples in the literature, the preemption is due to causal mechanisms that are triggered but fail. In our framework, this corresponds to *causal preemption*. Often, distinction is made between *early* and *late preemption*. E.g., in case of a lightening striking forest ground, the absence of trees (burnt by a previous fire, or cut down by a lumber company) is an early preempter of a forest

fire, while an extinction operation is a late preempter. In our framework, there is no general way to formally distinguish between early and late (causal) preemption, since the process semantics makes abstraction of the order of events.

Beside causal preemption, there is a second sort of preemption. Even when a causal mechanism r with effect L is applicable in world W, that is, all its conditions hold, it is possible that r does not fire. Intuitively, this corresponds to the situation when other causal mechanisms had caused L before r got the chance. We say that r is *chronologically preempted* in $\mathcal{P}$.

Example 4. *(***Window***, see (Hall 2004)) Suzy and Billy throw rocks at a window. Each throw is a separate causal mechanism causing the same deviant state of a broken window. We represent as follows:*

$$\left\{ \begin{array}{l} Broken \leftarrow SuzyT \,||\, \\ Broken \leftarrow BillyT \,||\, \end{array} \right\}$$

Assume that both throw, in which case the window will certainly break. In the overdetermination *scenario, they hit the window simultaneously. It corresponds to the causal process in which the fire set of $Broken$ contains both laws. In the* late preemption *scenario, Suzy's throw arrives first and smashes the window. It corresponds to the process in which only the first law belongs to the fire set of $Broken$. It is called here a case of* chronological preemption*. Observe that for the resulting world, this does not matter: the window is broken. Stated precisely, in the exogenous state $\{SuzyT, BillyT\}$, there are multiple possible causal processes. However, they are confluent: they lead to the same possible world.*

Adding firing information Several other sorts of knowledge have been claimed to influence our judgment of actual causes, e.g., whether a proposition is *normal* (like the presence of oxygen in the air), whether a proposition represents an intentional action of an agent, whether a causal mechanism fires. Such knowledge stands orthogonal to our approach; the language can be extended to express it. We illustrate this for knowledge about firing. We assume that causal laws in a theory Δ have a symbolic name, declared in expressions of the kind:

$$(BillyBreaks :) \quad Broken \leftarrow BillyT \,||$$

An extended causal theory is then a set of named causal laws and Boolean expressions ψ of symbols of Σ and atomic formulas $Fires(r)$ with r a name of one of the causal laws. Given a causal process $\mathcal{P}$, we define $\mathcal{P} \models \psi$ by the standard inductive rules for connectives and by the base rules that $\mathcal{P} \models p$ if $p \in World(\mathcal{P})$ and that $\mathcal{P} \models Fires(r)$ if r fires in $\mathcal{P}$.

Example 5. Window, late preemption, cont. *Suppose Billy's throw is chronologically preempted by Suzy's. In structural equations, it is frequent practice to encode this information by adding auxiliary variables and changing the structural equations (Halpern 2016b). But such knowledge is independent of the workings of causal mechanisms; it should better be expressed separately. In our logic, it amounts to knowledge that Billy's mechanism does not fire. It is expressed as:*

$$Broken \wedge \neg Fires(BillyBreaks)$$

The causal model extended with this proposition logically entails $Fires(SuzyBreaks)$.

(Fundamental) properties of causal knowledge

We first establish the link with structural equations. Recall that *predicate completion* (Clark 1978) transforms a causal theory Δ in a set $compl(\Delta)$ of structural equations.

Theorem 1. *The possible worlds of a causal theory Δ and the solutions of the structural equation model $compl(\Delta)$ are identical.*

The theorem gives an indication for the success of structural equations for causal reasoning even if they do not model informal key concepts of causation such as causal mechanisms and causal processes: the (many) problems that can be solved on the basis of the possible worlds of the theory (and of the variant theories obtained with interventions), can be solved using $compl(\Delta)$.

Proposition 1. *A causal process for Δ is uniquely determined by the set of its exogenous literals and firing rules. That is, two different processes differ on some exogenous literal or on the set of causal mechanisms that fire.*

Theorem 2. *Given a causal theory Δ, each exogenous state W_{exp} can be uniquely extended to a possible world of Δ. Thus, Δ is satisfiable in each exogenous state, and two different possible worlds of Δ differ on some exogenous literals.*

Theorem 3. *All causal processes of Δ in exogenous state W_{exo} realize the same world.*

The latter is a confluence theorem. It is one of these aspects that are brought to the surface by making the causal processes explicit. It tells something important about causal information. For a given exogenous state, it does not matter which of the rules are applied nor in what order they are applied: they will always result in the same world. This point was made in Vennekens, Denecker, and Bruynooghe (2009). A real world case would be that I send one friend the message that I won the lotto, and by the end of the day, I can be certain that all my friends know that I am rich. The process, the details of who tells who, may vary widely and is unknown to me; yet the outcome is predictable. It points to a valuable property of causal information: that it allows to derive much information about the state of the world that is the result of a causal process, even in the absence of almost any information on the process itself.

Proposition 2. *The causal language is non-monotonic: a world that is impossible in a causal theory Δ may be possible in an extension of Δ obtained by adding other causal laws to it.*

For a proof, consider the causal theory $\{ \; P \leftarrow Q \,||\, \}$. An impossible world is $\{P, \neg Q\}$. This world is possible after adding $P \leftarrow \neg Q \,||$. The original theory entails $\neg Q \Rightarrow \neg P$ while its extension does not.

Definitions of actual causation

A causal process $\mathcal{P}$ realizing world W provides a precise causal explanation of W from which different notions of causation can be "read off". Below it is used as a framework to define several notions of actual causation.

Definition 10. *A literal L is an influence of K in a possible causal process $\mathcal{P}$ of Δ if there is a path from K to L in $\mathcal{P}$.*

The concept of influence is useful but weak. We refine it to take the difference between triggering conditions and no-preemption conditions into account. When a causal mechanism fires and causes L, only its triggering conditions are seen as actual causes. E.g., when **Drinking poisoned coffee** without taking an antidote, drinking poisoned coffee is the actual cause of death, not the absence of antidote. Also, one cannot preempt a causal mechanism that has not been triggered, hence, when a causal mechanism to derive L remains inactive by a false triggering condition, its false no-preemption conditions are not actual causes of $\neg L$. E.g., when the victim takes the antidote but does not drink the poisoned coffee, the actual cause for survival is the absence of drinking, not the antidote. Only if the mechanism is active, will a false no-preemption condition be an actual cause of $\neg L$. E.g., the antidote is an actual cause of survival only if the victim drinks poisoned coffee.

Implementing these intuitions is easy: it suffices to discard all causal paths containing an arc $L \stackrel{r}{\rightarrow} K$ that is a no-preemption arc of a firing mechanism r or that is a preemption arc of a non-active causal mechanism r.

Definition 11. *A literal L is an actual P-cause of literal K in process $\mathcal{P}$ if there is a path $K \rightarrow \ldots \rightarrow L$ in $\mathcal{P}$ without no-preemption arcs and without preemption arcs of non-active causal mechanisms. Such path consists of trigger and no-trigger arcs, and preemption arcs of active causal mechanisms.*

The "P" stands for "preemption". Our both notions of causation are defined in the context of a causal process, whereas in most approaches actual causes are defined in the context of a possible world. We bridge this gap.

Definition 12. *A literal K is an influence (actual P-cause) of L in a possible world W of Δ if there is a possible causal process $\mathcal{P}$ realizing W such that K is an influence (actual P-cause) of L in $\mathcal{P}$. We call K a* definite *influence (actual P-cause) of L in W if it is an influence (actual P-cause) in every causal process realizing W. Otherwise it is called* speculative.

As pointed out by Vennekens (2011), even when we know the world, we may not know how it was caused and therefore, we may not be sure about the actual causes. This uncertainty is reflected in the above definition. It is illustrated by the different possible causal processes of **Window** in the exogenous context where both Suzy and Billy throw. It is one of these aspects that are brought to the surface by making the causal processes explicit.

Proposition 3. *The notions of influence and actual P-cause in processes and worlds are anti-symmetric and transitive.*

Now we turn to examples. The ones seen so far (**Arsenic and Strychnine, Drinking poisoning coffee** and **Window**) are modeled by simple causal theories having causal processes of depth 1. It is straightforward to derive the possible causal processes and the influences and actual P-causes of the endogenous literal. Moreover, as can be seen in the discussion preceding Definition 11, the results match the intuitions expressed in the introduction.

Example 6. *(***Backup (Hitchcock 2007) (early preemption versus switch)***)* A crime syndicate hires Assassin to poison victim's coffee who drinks it and dies. The syndicate had hired Backup to watch Assassin and to poison the victim in case Assassin would not poison the coffee. Backup did not have to intervene. *This scenario is a case of early preemption (of the poisoning by Backup). Three causal mechanisms can be discerned. They are represented:*

$$\left\{ \begin{array}{l} Dead \leftarrow APoison \,||\\ Dead \leftarrow BPoison \,||\\ BPoison \leftarrow \neg APoison \,|| \end{array} \right\}$$

The informal scenario corresponds to the context $\{APoison\}$, where the only actual P-cause of $Dead$ is $APoison$. This is the same answer as in (Bochman 2018) but certain counterfactual methods do not return $APoison$ as an actual cause (Lewis 1973; Halpern 2016a). In the context $\{\neg APoison\}$, the actual P-causes of $Dead$ are $BPoison$ and $\neg APoison$. That $\neg APoison$ is an actual P-cause is slightly disconcerting; perhaps this has to do with the longer length of the causal path from $\neg APoison$ to $Dead$. Still, we feel it makes sense, since the fact that Assassin does not poison, sets Backup's mechanism in motion to poison the victims coffee.

Now take an alternative story: the crime syndicate hires both Assassin and Backup, with a similar task: to pickup a poison at (the same) hidden place and poison victim. Assassin is ordered to go to the hiding place on Monday, Backup on Tuesday. The syndicate puts one potion of poison in the location on Sunday. *We argue that in this scenario, the causal laws are the same except for:*

$$BPoison \leftarrow || \neg APoison$$

Here backup has the plan to poison, but may be preempted to do so if Assassin took the poison. In the context $\{APoison\}$, the causes are identical as in the previous story. But in $\{\neg APoison\}$, only $BPoison$ is an actual P-cause and not $\neg APoison$. We feel this makes sense; after all, it was not $\neg APoison$ that triggered Backup to poison the victim, so how could it be a cause for $Dead$?

In both scenarios, $APoison$ is counterfactually irrelevant: whether true or false, the victim dies. The first is an early preemption scenario, the second is more like a **switch** *scenario, with $APoison$ as switch. It has been a challenge to explain the difference between early preemption and switch. This example suggests the underlying problem might be the distinction between triggering conditions and no-preemption conditions.*

Example 7. Double Preemption (Hall 2004) *Double preemption occurs when a potential preempter is preempted. It*

occurs in the following scenario. Suzy fires a missile (SF) to bomb target (B); enemy fires a missile (EF) to hit Suzy's missile (SMH) and Billy fires a missile (BF) to hit Enemy's missile (EMH). *We see three causal mechanisms:*

$$\left\{ \begin{array}{l} B \leftarrow SF \,||\, \neg SMH \\ SMH \leftarrow EF \,||\, \neg EMH \\ EMH \leftarrow BF \,|| \end{array} \right\}$$

In the causal process of context $\{SF, EF, BF\}$, the target is bombed. We find the causal path $BF \rightarrow EMH \rightarrow \neg SMH \rightarrow B$ which in the two last edges display a double preemption: the hit on enemy's missile preempts enemy's attempt at preempting Suzy's bombing.

We broaden the notion of actual P-cause to include double preemption. In our setting, a double preemption path is a causal path $K \stackrel{r}{\rightarrow} L_0 \rightarrow \ldots \rightarrow L_n \stackrel{r'}{\rightarrow} L$ ($n \geq 0$) such that

- $K \stackrel{r}{\rightarrow} L_0$ is a preemption arc of a causally preempted law r. In the example, it is the arc $EMH \rightarrow \neg SMH$.
- The n arcs $L_i \rightarrow L_{i+1}$ are arcs proving that L_0 is an actual P-cause of L_n (or identical to it). In the example, $n = 0$ and $L_0 = L_n$.
- $L_n \stackrel{r'}{\rightarrow} L$ is a no-preemption arc of a firing causal law r'. In the example, it is the arc $\neg SMH \rightarrow B$.

Definition 13. *The DP-causal graph of $\mathcal{P}$ consists of all arcs considered for actual P-causes augmented with* double preemption arcs $K \Rightarrow L$ *for every double-preemption path from K to L in $\mathcal{P}$. A literal K is an actual DP-cause of literal L in process $\mathcal{P}$ if there is a path from K to L in the DP-causal graph of $\mathcal{P}$.*

"DP" stands for double preemption. Billy's fire BF in **Double preemption** is a actual DP-cause of B although it is not an actual P-cause of B.

Proposition 4. *The actual P-causes are actual DP-causes; actual DP-causes are influences. The actual DP-cause relation is anti-symmetric and transitive.*

Example 8. *(***Triple preemption***) The new definition deals with triple preemption and more. Consider* **Double preemption** *extended with Jane who fires at Billy's missile (JF).*

$$\left\{ \begin{array}{l} B \leftarrow SF \,||\, \neg SMH \\ SMH \leftarrow EF \,||\, \neg EMH \\ EMH \leftarrow BF \,||\, \neg BMH \\ BMH \leftarrow JF \end{array} \right\}$$

In state $\{SF, EF, BF, JF\}$, the city is not bombed. The actual DP-causes of $\neg B$ are EF, BMH, JF. The DP-causal graph contains the double preemption arc $BMH \Rightarrow SMH$ which is induced by the double preemption path $BMH \rightarrow \neg EMH \rightarrow SMH$. Since Jane's fire JF is an actual P-cause of BMH, it is an actual DP-cause for the failed bombing $\neg B$.

Counterfactual dependence

In the first causal theory of **Backup**, Example 6, $APoison$ is an actual P-cause of $Dead$ in the exogeneous state $\{APoison\}$. We observed that $APoison$ was counterfactually irrelevant to $Dead$ in the sense that even if $APoison$ would have been false, $Dead$ would have been true all the same. In this section, we formally define this notion of counterfactual relevance.

Ever since the seminal work of Lewis (1973), actual causation has been analyzed using counterfactual reasoning. While our theory of actual causation does not rely on counterfactual reasoning, in many applications, counterfactual questions naturally arise. E.g., in **Window** "would the window have broken had Suzy not thrown?". Or, in **Backup**, "would the victim have died if Assassin had not poisoned him". Below, we define the notions of counterfactual (in)dependency and (ir)relevance using the concept of intervention, as introduced by Pearl (2000). We adapt the definition of intervention from (Vennekens, Denecker, and Bruynooghe 2010), where it was defined in the context of CP-logic.

Definition 14. *We define the* intervention *of causal mechanism r on causal theory Δ (denoted $\Delta[r]$) as the causal theory obtained from Δ by deleting all rules with the same head as r (if any) and adding r.*

Observe that the unique possible world of Δ extending exogeneous state W_{exo} is the unique possible world of $D + W_{exo}$ which is Δ extended with causal rules $L \leftarrow ||$ for each $L \in W_{exo}$.

Definition 15. *Let Δ be a causal theory without exogeneous symbols. Given that K, L are true in the unique possible world W of Δ, we define that L is counterfactually dependent on K according to Δ if L is false in the possible world of $\Delta[K \leftarrow \mathbf{f}]$.*

If Δ has exogeneous symbols and W is a possible world of Δ, we define that L is counterfactually dependent on K in W according to Δ if L is counterfactually dependent on K according to $\Delta + Exo(W)$.

If L counterfactually depends on K, we say also that K is counterfactually relevant for L.

Proposition 5. *If K is counterfactually relevant for L in W according to Δ then K is an influence of L in W.*

The inverse is not true, as can be seen in **Backup**.

Example 9. *(***Backup, cont.***) The causal theory corresponding to* **Backup** *in exogeneous state $W_{exo} = \{APoison\}$ is:*

$$\left\{ \begin{array}{l} Dead \leftarrow APoison \,|| \\ Dead \leftarrow BPoison \,|| \\ BPoison \leftarrow \neg APoison \,|| \\ APoison \leftarrow || \end{array} \right\}$$

According to this theory as well as the intervention by $\neg APoison \leftarrow ||$, $Dead$ is true. It follows that $APoison$ is counterfactually irrelevant for $Dead$. In terms of the original theory Δ, $APoison$ is a counterfactually irrelevant influence of $Dead$ in W_{exo}.

It can be seen that in context $\{SuzyT, BillyT\}$ of **Windows**, Suzy's throw is counterfactually irrelevant for $Broken$. Still, there is difference with the Assassins poisoning: if Suzy does not throw, then she does not actively contribute to breaking the window. But if Assassin does not poison, this causes the Backup to poison the victim, and so,

either way, Assassins choice is an actual cause of the victims death. The following definition expresses this stronger notion of irrelevance.

Definition 16. *Given Δ without exogeneous symbols, we define that K is a strongly irrelevant actual P-cause of L according to Δ if K is an actual P-cause of L in the world of Δ and $\neg K$ is an actual P-cause of L in the world of $\Delta[K \leftarrow \mathbf{f}]$.*

In general, if Δ has exogeneous symbols and W is a possible world of Δ, we define that K is a strongly irrelevant actual P-cause of L in W according to Δ if K is a strongly irrelevant actual P-cause of L according to $\Delta + Exo(W)$.

We see that $APoison$ is a strongly irrelevant actual P-cause of $Dead$ while $SuzyT$ is counterfactually irrelevant but not a strongly irrelevant actual P-cause of $Broken$.

We have a strong intuition not to consider strongly irrelevant actual P-causes such as $APoison$ as actual causes whereas according to our intuition, other actual P-causes such as $SuzyT$ are actual causes even if they are counterfactually irrelevant.

We argue that it is an asset of our approach that it helps distinguishing between *inherently* counterfactual questions about actual causation (such as "would the window have broken if Suzy had not thrown"), and the use of counterfactual analysis to exhume actual causes from a modelling that does not express the causal process.

Related work and conclusions

In the spectrum of counterfactual versus regularity-theoretic approaches to actual causation, our method belongs to the second category since it is based on analysis of causation in the actual world and the actual causal process.

Counterfactual methods originated from Lewis' idea of interpreting "C caused E" as the statement "without C, E would not have been". When counterexamples kept emerging, ever more sophisticated counterfactual strategies were developed. Present-day methods derive actual causes from "blackbox" theories in a way that seems to mimic an empirical scientist who, perhaps without knowledge of the causal mechanisms in the domain, tries to discover actual causes by a strategy of experiments in which the values of well-chosen variables are varied.

Our definition of actual causes defines actual causes in terms of the dependencies shown in the actual process. That leads to a very different definition than the counterfactual definitions of actual causation. Still we do not think that they contradict with each other, but rather than that they point at complementary aspects of the same thing.

Actual causation is an informal concept, which each of us learns through experience and communication with other people. It is not a concept that we acquire by receiving a definition. In such cases, it is well possible that two very different definitions of a concept cover to a large extend the same set of phenomena. We think this is the case with the counterfactual definitions of actual causation, and the dependency-based definitions that we used. In fact, there is a clue to this in the very concept of "dependency": if A depends on B, then we expect that if B is not, then A might not be, and this suggests a counterfactual dependency.

It should not be easy to reconstruct the exact factual dependencies using counterfactual experiments. Sometimes, "nature" puts an effort to hide certain dependencies, and no experiment can bring a dependency to the surface. It suggests that counterfactual definitions and dependency definitions do not perfectly match. Nevertheless, we expect this to be an exception; we expect that in many cases, counterfactual experiments are able to bring a dependency to the surface. This raises a research question: what is the correspondence between actual causes in causal theories Δ defined in terms of the actual causal process, and actual causes derived through counterfactual methods from $comp(\Delta)$. This is a topic for future work.

In this framework, we studied several sorts of knowledge that are important for actual causation but are not or not well expressed in many causal languages: knowledge of causal mechanisms, triggering versus preempting conditions, and whether they fire. We proposed a causal logic suitable for modular expression of such knowledge and equipped with a possible causal process semantics. The explicit modeling of causal processes brought a few fundamental aspects of causation to the surface: e.g., the convergence of causal processes and, paradoxically, theorems explaining why many useful causation problems can be solved without modelling mechanisms and processes. Using causal processes as an explanation of the world, we provided definitions for several notions of actual causation including double preemption. Further analysis is required to corroborate and refine these results, but the method handles a range of problematic examples in causal reasoning.

The aim to study actual causation in the context of causal processes is present in neuron diagrams approaches (Lewis 1986). However, neuron diagrams do not represent individual causal mechanisms (similar to a structural equation) and do not distinguish between triggering and preempting conditions, and fall short for the sort of examples that motivated this paper. The first causal reasoning study in a language that accounts for causal mechanisms, processes and worlds was (Vennekens, Denecker, and Bruynooghe 2009). The language CP-logic was used for different forms of reasoning such as probabilistic reasoning, interventions and actual causation. The logic defined here is related in spirit to CP-logic but differs from it quite considerably. E.g., causal processes are formalized differently, and several sorts of knowledge studied here cannot be expressed in CP-logic (and vice versa). The actual causation method for CP-logic proposed by Vennekens (2011) and refined by Beckers and Vennekens (2012) is based on causal processes as well, but it is intuitively and mathematically completely different. It is a counterfactual method based on analysis of alternative causal processes, in a way related to the approaches of Hall (2004; 2007). The relation with our approach is not obvious and we leave a further analysis of this for future work.

Our formalism is simple and propositional. To make it suitable to express real-world causal domains, it needs to be extended to the predicate case, with quantification, formulas in the body, the possibility to define auxiliary concepts, non-deterministic causation, probabilities, cyclic causation, etc. such that definitions of actual causation still work. This

is for future work. CP-logic covers already most of these extensions, so we expect this to be feasible.

Implementation

We specified the different notions of causality of this paper as a first order logic theory using the knowledge base system IDP (De Cat et al. 2016). Our model is available at `http://adams.cs.kuleuven.be/idp/server.html?chapter=intro/11-AC`. By applying model expansion inference on this specification and on structures encoding causal theories, various notions of actual causation are computed. The webpage contains all examples of the paper and several others. Readers can modify these examples or edit their own and run the system in the web browser to solve causal questions.

Acknowledgements

We are grateful to Alexander Bochman and Sander Beckers for many discussions and valuable feedback. Bart Bogaerts is a postdoctoral fellow of the Research Foundation – Flanders (FWO).

References

Baumgartner, M. 2013. A regularity theoretic approach to actual causation. *Erkenn* 78(Suppl 1:85.

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 Starting AI Researchers Symposium, STAIRS, Montpellier, 27-28 August 2012*, volume 241, 35–46.

Bochman, A. 2018. Actual causality in a logical setting. In *IJCAI*.

Clark, K. L. 1978. Negation as failure. In *Logic and Data Bases*, 293–322. Plenum Press.

De Cat, B.; Bogaerts, B.; Bruynooghe, M.; Janssens, G.; and Denecker, M. 2016. Predicate logic as a modelling language: The IDP system. *CoRR* abs/1401.6312v2.

Fenton-Glynn, L. 2015. A proposed probabilistic extension of the halpern and pearl definition of 'actual cause'. *The British Journal for the Philosophy of Science*.

Gerstenberg, T.; Goodman, N. D.; Lagnado, D. A.; and Tenenbaum, J. B. 2015. How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 782–787.

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* 175(2):169–192.

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

Halpern, J. 2016a. *Actual causality*. MIT Press.

Halpern, J. Y. 2016b. Appropriate causal models and the stability of causation. *Rew. Symb. Logic* 9(1):76–102.

Hiddleston, E. 2005. A causal theory of counterfactuals. *Noûs* 39(4):632–657.

Hitchcock, C. 2007. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review* 116(4):495–532.

Hume, D. 1739. *A Treatise of Human Nature*. John Noon.

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

Lewis, D. 1986. Postscripts to 'causation'. In Lewis, D., ed., *Philosophical Papers Vol. Ii*. Oxford University Press.

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

Vennekens, J.; Denecker, M.; and Bruynooghe, M. 2009. CP-logic: A language of causal probabilistic events and its relation to logic programming. *TPLP* 9(3):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.