Explaining actual causation in terms of possible causal processes*

Marc Denecker[1], Bart Bogaerts[2,1], and Joost Vennekens[3,1]

[1] KU Leuven, Department of Computer Science
[2] Vrije Universiteit Brussel (VUB), Department of Computer Science
[3] KU Leuven, Department of Computer Science, Campus De Nayer
{firstname.lastname}@cs.kuleuven.be

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 what conditions can make them fail.
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 framework for actual causation in which various "production" notions of actual causation are defined. Contrary to counterfactual definitions, these notions are defined directly in terms of the (formal) causal process that causes the possible world.

1 Introduction

Since the days of Hume [21], causal reasoning has been an active research domain in philosophy and (later) knowledge representation. With the groundbreaking work of Lewis [22] and Pearl [25], the structural equations and counterfactual reasoning approach became mainstream [15, 14, 9, 10]. But the debate remains intense [11]. The counterfactual approach is contested by some [13, 1, 5]. In many scenarios, there is no agreement of what the actual causes are, 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; it also show that many sorts of knowledge influence our judgment of actual causation. Science is not ready yet with unraveling all this.

Among 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. Indeed, such examples are particularly interesting, because they demonstrate that some relevant knowledge is missing from the causal model. A powerful illustration is given by Halpern [16], who discusses 6 causal examples from the literature in which authors showed (often convincingly) that the actual causation definition of Halpern and Pearl [15], henceforth called HP, failed to

* Bart Bogaerts is a postdoctoral fellow of the Research Foundation – Flanders (FWO).

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**" and "But all this talk of mechanisms [...] suggests that the mechanism should be part of the model."

That is, he argues that we should make knowledge of causal mechanisms explicit.

That such information is relevant for causal reasoning is not surprising. Many causal scenarios in the literature comes with an informal specification of causal mechanisms and, often, a sometimes partial *story* specifying which mechanisms are active and how they are rigged together in a causal process. As observed before [11, 27], most of this information is abstracted away in structural equation models. We illustrate to what problems this may lead with a simple example, an *ambiguity* of the same sort as tackled by Halpern [16]. Consider two scenarios involving two deadly poisons, arsenic and strychnine. In the first scenario, intake of any of these poisons triggers a deadly biochemical process. The corresponding structural equation is

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

If both poisons are taken, this is an instance of overdetermination; HP derives 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* (i.e., in both cases, there are 4 possible worlds: $\{D, A, S\}$, $\{D, A, \neg S\}$, $\{D, \neg A, S\}$ and $\{\neg D, \neg A, \neg S\}$)! 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. What is missing is more detailed information about the *causal processes* that generate the possible worlds and about the individual causal mechanisms that constitute these processes.[4]

The following scenario, simplified from **Assassin** [20], illustrates another relevant sort of knowledge that is not expressed in structural equation models.

[4] A more intuitive structural equation for the second scenario is $Arsenic \vee (\neg Arsenic \wedge Strychnine)$. It is equivalent under standard semantics to the original equation. Nevertheless, it suggests an alternative way to resolve the ambiguity: developing a more refined semantics that distinguishes between the two equations. We suspect that structural equations under such a refined semantics might turn out to be quite similar to the logic we develop in this paper.

An assassin may kill a victim by administering deadly poison. A bodyguard may rescue the victim by administering an antidote. The structural equation:

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

correctly characterizes the possible worlds. However, 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 argue below, this distinction plays a role in many controversial causal examples and should be added to the causal model.

Halpern's solution to the first type of ambiguities is to *reify* the different causal mechanisms by auxiliary variables and structural equations representing when the mechanism *fires*. He applies this methodology to explicate the causal mechanisms in the different interpretations of each of the 6 cases. The causal models of the refined theories then not only encode the actual world, but also (part of) the causal process that creates it. For all 2×6 cases, HP was able to detect the intuitively expected actual causes using the refined theories.

These are great results, but they also raise some fundamental questions. First, Halpern's approach is to refine existing structural equation models to resolve *reported* ambiguities on them. What guarantee is there that all ambiguities are resolved now? To eradicate the problem of such ambiguities, we need a modelling language that supports expression of individual causal mechanisms. This is the first topic on which our paper contributes. Second, his analysis shows that knowledge of individual causal mechanisms and which of them fire influences our judgment of actual causation. But this does not explain how this works. Sure, HP was powerful enough to produce the expected answers, but HP is not based on causal mechanisms and processes, hence this method cannot explain why and how causal processes determine the actual causes. What is missing is a principled explanation of actual causation in terms of the causal process and the causal mechanisms. This is the second topic on which our paper contributes. Third, the second ambiguity, the one that appears in **Assassin**, is not a problem of discerning different causal mechanisms and Halpern's methodology is not applicable to this case. We argue that to resolve this type of ambiguities, it is necessary to express the distinction between conditions that *trigger* the causal process and conditions that, if false, *preempt* the causal mechanism. We propose a modelling language for this and we argue that making this distinction explains a number of controversies in causal reasoning, such as the difference between *early preemption* and *switch* scenarios.

2 The causal logic: syntax and informal semantics

We propose a propositional causal modelling language to resolve the reported ambiguities. It can be lifted easily to the predicate level but this would merely increase the formal complexity without contributing to the essence of the paper.

To represent a causal domain, a *vocabulary* Σ of propositional symbols is chosen; each symbol expresses an atomic proposition in the domain. Literals 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 $L = P$. As usual, we distinguish between *endogenous symbols*, for which the mechanisms that cause them are expressed in the theory, and *exogenous* symbols, for which no causal mechanisms are expressed.

A causal theory is a set of causal mechanisms. Each mechanism has *triggering conditions*, which set the mechanism in operation; *enabling conditions*, which if false, preempt the mechanism; and an *effect*. This leads to the following definition.

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

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

where
- $\leftarrow$ *is the causal operator (not material implication),*
- L *is a literal of an endogenous symbol, called the* effect*,*
- T *is a sequence of literals called* triggering conditions*,*
- C *is a sequence of literals called* enabling conditions*.*

The causal mechanism $L \leftarrow ||$ *represents the unconditional causal mechanism causing* L*. Elements of* $T \cup C$ *are called* conditions *of the causal mechanism.*

A causal theory Δ *is a set of causal mechanisms that contains at least one mechanism for each endogenous symbol and such that:*
- Δ *is acyclic, i.e., there exists a strict well-founded order on symbols such that for each causal mechanism, the symbol in the effect is strictly larger than the symbols of the conditions.*
- Δ *does not contain mechanisms with contradictory effects* $P \leftarrow \ldots$ *and* $\neg P \leftarrow \ldots$*.*

The logic imposes two main constraints on causal theories: acyclicity and absence of constradictory effects. In many causal domains, cycles in causal mechanisms exist. Cycles are allowed in several causal rule formalisms [28, 6, 5, 7]. Following [28], the logic proposed here can easily be extended with cycles. We do not implement it since the greater complexity would detract attention from the essence of this paper: the resolution of ambiguities. The absence of contradictory effects is an inherent aspect of the language and is explained below.

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

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

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

x respectively. Three of these rules have the empty sequence of enabling conditions. In the last rule, strychnine poisoning is triggered by strychnine but preempted by arsenic.

We now describe the informal semantics of the language. A causal theory does not impose constraints on the exogenous symbols and makes abstraction of causal mechanisms affecting them. For the endogenous symbols, the causal theory is assumed to contain *all* causal mechanisms affecting them. Each endogenous proposition has a *default* state L and a *deviant* state $\neg L$. The effect of a causal mechanism is always the deviant literal. A causal process starts in a state where endogeneous properties are in their default state, and proceeds by firing *applicable but unsatisfied causal* mechanisms: mechanisms with true conditions but false effect. Firing a causal mechanism switches the effect on, moving the included proposition from its default to its deviant state. Once a deviant literal is true, it remains true. As such, with each endogenous proposition zero or one *event* is associated: zero if it stays in its default state, one if it switches. Such a switch event may be caused by multiple causal mechanisms causing L simultaneously. The process stops when all applicable causal mechanisms are satisfied. The resulting state is a possible causal world of the theory.

It can be seen that a form of the *law of inertia* is present in the logic: an endogenous symbol remains in the same state unless it is affected by some causal mechanism. Also, endogenous properties have a fixed default and deviant state and causal mechanisms cause deviant literals; hence, mechanisms have no contradictory effects.

The causal processes considered here are clearly of a limited kind. In many causal domains, endogenous properties evolve from true to false and back again, caused by mechanisms with contradictory effects. E.g., flipping a switch causes the light to be on if the light is not on and vice versa. Such domains, interesting as they are, fall outside the scope of this paper. First of all, the causal ambiguities studied here arise in causal domains modelled in non-temporal causal languages (structural equations, causal neuron diagrams, causal calculus, CP-logic, ...). We argue that in the majority of such applications, the causal processes are of the simple kind considered here. Also, building a language for modelling causal worlds that are the result of complex dynamic causal processes is conceptually, mathematically, and computationally complex (e.g., causal processes may not terminate). This is outside the scope of this paper.

In worlds caused by causal processes of the sort we consider here, every deviant literal L has a causal explanation, namely the causal mechanism(s) that caused it. On the other hand, a default literal L that holds in the world is not caused by a specific causal mechanism; it is true by inertia. Nevertheless it has a causal explanation as well, namely, that every causal mechanism for $\neg L$ is blocked. Either way, the language implements Leibniz's principle of sufficient reason —that every true fact has a reason— (but only for endogenous facts).

Definition 2. *A* world W *is a complete and consistent set of literals, i.e., a set of literals such that for each* $P \in \Sigma$*, 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* $Exo(W)$*. As usual, an exogenous state is called a* context*. A symbol* P *is in its* deviant state *in* W *if its deviant literal holds in* W*, and in its* default state *otherwise.*

Definition 3. *A causal mechanism* r *of the form* $L \leftarrow A \,||\, B$ *is* blocked *by a condition* $K \in A \cup B$ *in world* W *if* $\neg K \in W$*. The mechanism* r *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 mechanism is* applicable *in* W *if* $A \cup B \subseteq W$*. The mechanism* r fails *in* W *if it is active but is blocked by an enabling condition in* W*. A causal mechanism is* satisfied *in* W *if it is blocked, or if its effect holds in* W*.*

Triggering conditions versus enabling conditions The distinction between triggering and enabling conditions of causal mechanisms is a new feature of our language. 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 two conditions are needed: a spark igniting a hotbed in the forest and absence of extinction operations. There is no difference between the two conditions on the level of counterfactual dependence. Nevertheless, it is the spark (in the form of a lightning or an unsafe camp fire) that triggers the causal mechanism; the condition of absence of extinction operations is there only because such operations would make the causal mechanism fail in achieving its effect. 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 absence of fire extinction. Our goal here is to propose a formalization of this notion of actual causation. To define it, the nature of the conditions must be clear from the causal theory.

Example 2. (Hitchcock's **Assassin**, [20, p. 504]) Drinking coffee poisoned by Assassin causes Victim to die unless an antidote is administered by Bodyguard. We discuss three conditions here: presence of poison in the coffee ($Poison$), drinking the coffee ($Drink$), and absence of antidote ($\neg Antidote$). The poisoning process is physically triggered by the event of drinking the coffee. However, it is the intake of poison that triggers the poisoning process. Thus, the triggering conditions are $Drink$ and $Poison$. Intake of an antidote causes the process to fail in achieving its effect. So, we argue for the following representation:

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

Hitchcock pointed at the different "strengths" of the first conditions versus the third condition as causes for death. When the three conditions are true and Victim dies, drinking poisoned coffee seems to be a "stronger" cause of this than the absence of antidote. He argues that this is because absence of antidote is an *omission*, in particular, of the event of administering antidote. However, this cannot be the explanation. First, omissions are frequently perceived as strong causes [26]. Second (and illustrating the first point), if there is no poisoning or

no drinking but antidote is administered, the omission of one or both of the first conditions seems to be the "stronger" actual cause for survival than the presence of antidote. The explanation we propose is in terms of triggering versus enabling conditions. When triggering and enabling conditions are true, we perceive the triggering conditions as "stronger" actual causes for the deviant effect than the enabling conditions. When triggering conditions and enabling conditions are both false, we see the omission of triggering conditions as the "stronger" actual causes for the absence of the effect; after all, if the causal mechanism is not even triggered, the falsity of its enabling conditions does not seem to matter. The only situation where an enabling condition plays a role as an actual cause is when the mechanism is active (its triggering conditions hold) but fails due to falsity of the enabling condition.

Even now, before having defined a formal semantics, it is intuitively clear how to transform causal theories to structural equations, namely by *predicate completion* [8]. 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 and enabling conditions.

3 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 et al. [28] formalize them as sequences of states in which at every state one causal mechanism is applied until all causal mechanisms are satisfied. This representation is precise and gives an account of, e.g., the "stories" in many causal examples. However, its high level of detail is not actually required for dealing with actual causation, 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 mechanisms. Let Δ be a causal theory throughout the rest of the paper.

Definition 4. *A possible causal process for Δ is a directed labeled graph $\mathcal{P}$ whose set of nodes is a world, denoted $World(\mathcal{P})$. Each arc from literal K to literal L is labeled with a mechanism r or $\neg r$. 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 mechanisms with head L, called the* firing set *of L, such that for each condition K of each $r \in F_L$, there is an arc $L \overset{r}{\leftarrow} K$. There are no other arcs to L.*

- *For each default endogenous literal $L \in World(\mathcal{P})$, for each mechanism $r = \neg L \leftarrow \ldots$, the set B_r of conditions of r that are false in $World(\mathcal{P})$ is non-empty and there is an arc $L \overset{\neg r}{\leftarrow} \neg K \in \mathcal{P}$ for each $K \in B_r$. There are no other arcs to L.*

When the causal process in the real world leading up to the current world can be observed, the corresponding formal process can be extracted along the following lines. At each time the state changes during the process, one or more deviant literals L become true. For each, one detects the mechanisms that caused it and adds arcs from its conditions to L. When a default literal L remains true, one investigates why the mechanisms that could cause $\neg L$ did not fire, and adds arcs from the negation of all their false conditions to L.

Definition 5. *We call arcs $L \overset{r}{\leftarrow} K$* active arcs *and distinguish between* trigger arcs *and* enabling arcs *depending on the type of the condition K is in r. We call arcs $L \overset{\neg r}{\leftarrow} \neg K \in \mathcal{P}$* blocking arcs *and we distinguish between* nontrigger arcs *and* failure arcs *depending on the type of condition K is in r.*

The causal process semantics induces a possible world semantics.

Definition 6. *A causal process $\mathcal{P}$* realizes *the world $World(\mathcal{P})$. We call W a possible world of causal theory Δ if it is realized by some causal process for Δ.*

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

Definition 4 treats triggering and enabling conditions symmetrically, except for the names of arcs. As a consequence, the classification of the conditions in causal mechanisms has no impact on the possible causal worlds. However, it will play a key role in the definition(s) of actual causation.

Proposition 1. *A world W is a possible causal world of causal theory Δ iff W is a model of the completion of Δ.*

As such, the extra information[5] available in a causal theory Δ compared to its corresponding structural model (i.e., its completion) does not affect the possible worlds nor does it affect the answer to any inference problem that can be resolved by reasoning on possible worlds. However, it does affect the actual causation question. This was shown by the ambiguities.

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 mechanism is active and fires and $\neg Alive$ has incoming trigger arcs from $Poison$ and $Drink$ and an enabling arc from $\neg Antidote$. In context $\{Drink, Poison, Antidote\}$, the mechanism is active but fails; $Alive$ has an incoming failure arc from $Antidote$. In $\{\neg Drink, \neg Poison, Antidote\}$, the mechanism is inactive and $Alive$ has nontrigger arcs from $\neg Drink, \neg Poison$ and a failure arc from $Antidote$. The latter context corresponds to **Bogus Prevention** [18, 12].

[5] Information on different mechanisms, and triggering conditions versus enabling conditions.

The firing set F_L of a deviant literal L may contain more than one mechanism, in which case L is *overdetermined*.

In our framework, three sorts of preemption of causal mechanisms can be distinguished. The first one is that the causal mechanism is blocked by some triggering condition and is inactive. The second is that the mechanism is active but blocked by an enabling condition and thus *fails*. The third sort of preemption occurs when a causal mechanism r with effect L is applicable in world W (all its conditions hold) but L was caused by other mechanisms. This corresponds to *late preemption*.

Example 4. (**Window**, see [13]) Suzy and Billy throw rocks at a window. Each throw is a separate causal mechanism causing the same deviant state.

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

Assume that both throw. 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 $Broken$. 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.

4 Definitions of actual causation

The informal notion of actual causation is vague and overloaded with many different intuitions. It is the role of science to unravel these. Below, we propose several distinguished notions in the context of possible process semantics. A causal process $\mathcal{P}$ realizing world W provides a precise causal explanation of W. We define several notions of causation that can be "read off" from the actual causal process. They are *objective* notions in the sense that they are defined in terms of the *actual* causal process that shaped the actual world. In this respect, they are similar to the notion of *production* in [13] in the context of causal neuron diagrams, and they can be contrasted with counterfactual notions of causation such as HP which are defined in terms of a class of *hypothetical* worlds. Generally stated, we interpret a notion of actual causation as a "production" notion if it can be derived from the actual causal process only. The different notions of causation below are defined in terms of different sorts of causal paths in the causal process.

Definition 7. *Given a causal process $\mathcal{P}$ for a causal theory Δ, a literal K is an influence of literal L in $\mathcal{P}$ if there is a path from K to L in $\mathcal{P}$.*

An influence of a literal L in a process $\mathcal{P}$ is any fact that has influenced the causal process causing L. We view the notion of influence as a "lower bound" for

causation: in any reasonable notion of actual causation, an actual cause should be at least an influence. Stricter notions can be defined by limiting the paths that are considered: for instance, we call an influence *active* if the path contains only active arcs, i.e., if there is a chain of firing mechanisms between K and L. Active influences are similar to the notion of actual causation defined in [17] in the context of neuron diagrams.

The notion of influence is defined in terms of the causal process, whereas in most approaches actual causes are defined in the context of a possible world. As pointed out by Vennekens [27], 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 emerged in the different possible causal processes of **Window** when both Suzy and Billy throw. This uncertainty is reflected in the definition below.

Definition 8. *A literal K is a* possible influence *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 of L in $\mathcal{P}$. We call K a* definite influence *of L in W if it is an influence in every causal process realizing W.*

The notion of influence does not distinguish between triggers and enabling conditions. As we argued in the introduction and in Example 2, this sometimes leads to counterintuitive results. We now examine how the distinction between triggers and enabling conditions affects our judgment of causation. This will lead us to further refine the notion of influence. We consider three variants of the poisoned coffee example. **(1)** Victim drinks poisoned coffee without having received the antidote. In this case, $\neg Antidote$ is an influence of $\neg Alive$. Yet, the intuition, originally expressed by Hitchcock, is that it is the poison that caused his death, not the absence of antidote. In general, when L is caused by a mechanism r, the triggering conditions of r are actual causes for L, while its enabling conditions are not. **(2)** Victim is given an antidote, but the coffee is not poisoned. Here, both $\neg Poison$ and $Antidote$ are influences. However, only the absence of poison should be counted as an actual cause for his survival, not the antidote, since one cannot preempt an inactive mechanism. In general, if a mechanism r for $\neg L$ is inactive, the actual causes of L are the false triggering conditions of r, not its false enabling conditions. **(3)** Victim is poisoned and receives an antidote. Here, the antidote is an influence, and it *is* a cause for his survival. In general, if r fails, its false enabling conditions are causes.

These intuitions are implemented in the following definition.

Definition 9. *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 enabling arcs and without failure arcs of non-active causal mechanisms. Such a path consists of trigger and nontrigger arcs, and failure arcs of active causal mechanisms. We say that L is a* direct actual P-cause *of K if the length of the path is 1, and an* indirect actual P-cause *otherwise.*

The "P" stands for "production", the basic "material" sort of causation of this causal language, similar to production in [13, 3]. This concept can be further constrained, e.g., to the notion of *active* actual P-cause.

The notion of actual P-cause is extended from processes to worlds in exactly the same way as was done for the notion of influence in Definition 7.

Proposition 2. *The notions of influence and actual P-cause in processes and worlds are transitive.*

All examples seen so far (**Arsenic and Strychnine, Drinking poisoning coffee** and **Window**) are modelled by simple causal theories which in every context has causal processes of length 1. As can be seen in the discussion preceding Definition 8, the actual P-causes of the endogenous literal match the intuitions expressed in the introduction.

Example 5. Assume in **Assassin**, that a crime syndicate ordered the murder. In this scenario, the following causal mechanism is in operation.

$$Poison \leftarrow CS\_Order$$

In the context $\{CS\_Order, Drink, \neg Antidote\}$, $Poison$ is a direct actual P-cause of $\neg Alive$ while CS_Order is an indirect actual P-cause of $\neg Alive$.

Example 6. **Double Preemption [13]** 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 (annotated with names r_1, r_2, r_3).

$$\left\{ \begin{array}{ll} B \leftarrow SF \,||\, \neg SMH & (r_1) \\ SMH \leftarrow EF \,||\, \neg EMH & (r_2) \\ EMH \leftarrow BF \,|| & (r_3) \end{array} \right\}$$

In Figure 1, three causal processes are graphically displayed. Red nodes are deviants, green nodes are defaults and grey nodes are exogenous. Full black arcs leave from trigger conditions; dotted purple arcs from enabling conditions. The arc is active if it ends in a red deviant node, it is blocking if it ends in a green default node. A third type of green arc leaves from a trigger condition in an active but failing mechanism that is preempted by an enabling condition.

The left process shows the causal process of context $\{SF, EF, \neg BF\}$ where Suzy's missile is destroyed by enemy fire and target is not bombed. The actual P-causes of $\neg B$ are SMH, EF. The middle shows the process in context $\{\neg SF, EF, \neg BF\}$ where the actual P-cause for $\neg B$ is false trigger $\neg SF$ but where SMH is another influence of $\neg B$. The right shows the causal process of context $\{SF, EF, BF\}$, where everyone fires, enemy's antimissile is destroyed by Billy's and the target is bombed. The actual P-cause of B is SF while $\neg SMH, EMH, BF$ are influences. The causal path $BF \rightarrow EMH \rightarrow \neg SMH \rightarrow B$ shows in the two last edges display a double preemption: the hit on enemy's missile preempts enemy's attempt at preempting Suzy's bombing. Some view that Billy's fire BF is an actual cause of B by double preemption. While this is not derived in our definition of actual P-cause, this kind of pattern can be read off from the causal process and is not difficult to formally define in the framework.

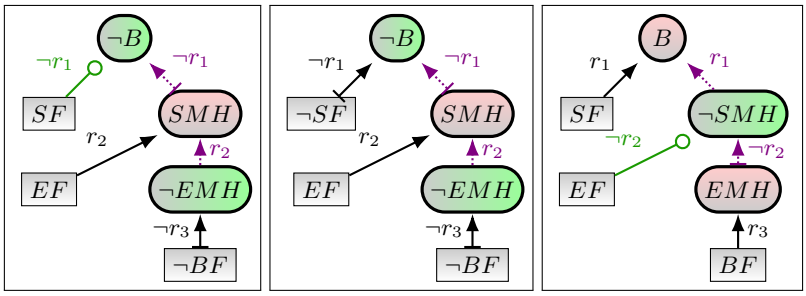

Fig. 1: *Graphical representation of three causal processes.*

4.1 Early preemption versus switch

A well-known issue in the actual causation literature concerns the relation between *Early Preemption* and *Switching* examples. Let us illustrate this by means of the following example of Early Preemption.

*Example 7 (***Backup [20]***).* An assassin-in-training is on his first mission. Trainee is an excellent shot: if he shoots, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim. The following is the standard structural equation model used in the literature for this story, where the context is such that $Trainee$ is true.

$$Victim := Trainee \vee Supervisor$$
$$Supervisor := \neg Trainee$$

A standard example of *Switching* is the following example.

*Example 8 (**Dog Bite** [24]).* Terrorist, who is right-handed, must push a detonator button at noon to set off a bomb. Shortly before noon, she is bitten by a dog on her right hand. Unable to use her right hand, she pushes the detonator with her left hand at noon. The bomb duly explodes. A standard structural equation model of this example is as follows, where the context is $\{Bite\}$.

$$Bomb := LH \vee RH$$
$$LH := Bite$$
$$RH := \neg Bite$$

Let us now compare these two examples. The role of $Trainee$ and $Bite$ in the formal models of both examples are remarkably similar. Nevertheless, the common opinion is that in cases of Early Preemption, there *is* causation (i.e., Trainee caused Victim's death), whereas in cases of Switching, there is no

causation (i.e., dog's bite did not cause the bomb to go off, even though it did cause Terrorist to push the detonator with her left hand).

The similarity between both causal models becomes even more striking when we extend the first example by an intermediate variable *Bullet*, that represents the fact that a bullet leaves Trainee's gun. In this case, we obtain:

$$\begin{aligned} Victim &:= Bullet \vee Supervisor \\ Bullet &:= Trainee \\ Supervisor &:= \neg Trainee \end{aligned}$$

This model is now formally identical to the model for Dog Bite. However, the addition of the intermediate variable *Bullet* seems like it should not affect our causal judgments. Therefore, we now essentially have one formal model, from which we nevertheless would expect different actual causation answers depending on the informal interpretation. This is an ambiguity.

In the literature, several solutions have been proposed for this problem. According to Weslake [29], it is precisely the addition of the intermediate variable *Bullet* that turns this example from an instance of Early Preemption into an instance of Switching and he therefore does not agree with our intuition that adding this variable should preserve our causal judgments. Becker and Vennekens [4] and Hall [12] argue that an action such as shooting a victim can never be completely deterministic and that it is therefore necessary to change to a non-deterministic (or probabilistic) representation in order to correctly handle the examples. Hitchcock [19] and Halpern and Pearl [15] argue that the formal model we presented is not suitable for Switching examples, based on an analysis of the counterfactual interventions admitted by the model.

We offer a different explanation which we believe goes more to the heart of the difference between the two kinds of examples. Our opinion is that the ambiguity is of the same type as the second ambiguity in the introduction, the ambiguity in the context of **Assassin**: the difference between early preemption and switching is located in the subtle but important difference between triggering and enabling conditions.

If we now re-examine the two above examples, we see that, first of all, $Trainee$ is obviously a triggering condition for *Bullet*, since it is Trainee's pulling of the literal trigger that sets in motion the causal mechanism that leads to the bullet's exiting the gun. The relevant causal mechanisms are therefore:

$$\left\{ \begin{array}{l} Victim \leftarrow Bullet \,||\\ Victim \leftarrow Supervisor \,||\\ Bullet \leftarrow Trainee \,||\\ Supervisor \leftarrow \neg Trainee \,|| \end{array} \right\}$$

However, analyzing the Dog Bite example, we notice that neither the dog bite nor its absence is actually a trigger for the causal mechanism that leads to the detonator being pushed. Indeed, for all we know, the dog bite could have happened (or failed to happen) a long time before the detonator was actually

pushed, giving Terrorist plenty of time to have a change of heart in between. The real trigger for the mechanism is Terrorist deciding to detonate the bomb. In our language, this example is therefore more appropriately modelled as:

$$\left\{\begin{array}{l} Bomb \leftarrow LH \,||\\ Bomb \leftarrow RH \,||\\ LH \leftarrow DecideToDetonate \,||\, Bite\\ RH \leftarrow DecideToDetonate \,||\, \neg Bite \end{array}\right\}$$

(Here, we introduced the new variable $DecideToDenote$ for clarity, but nothing changes in our analysis if we do not do this and leave the set of triggers empty.)

By making the distinction between triggering conditions and enabling conditions, we believe to have a convincing answer to the question of what really distinguishes early preemption from switching. Indeed, our definitions now yield that $Trainee$ is a actual P-cause of $Dead$, while $Bite$ is not.

5 Related work and conclusions

We studied several sorts of knowledge that are important for actual causation: knowledge of causal mechanisms and which of them fire, and the distinction between triggering and enabling conditions. Causal mechanisms with enabling conditions can be considered as mechanisms with a failure option: a false enabling condition leads to failure of the mechanism. The relevance of these concepts was brought to light by ambiguities. We proposed a language to express them and defined a possible causal process semantics, which induces a possible world semantics. Using causal processes as an explanation of the world, we provided definitions for several notions of actual "production" causation. The notion of (active) influence is independent of the distinction between triggering and enabling conditions, while (active) actual P-cause takes them into account. We argued that the distinction explains the difference between early preemption and switching.

We evaluated these ideas in a range of examples. Our test set includes those of [11], where the definitions of Woodward [30] and HP [15] are put to the test. The notion of actual P-cause correctly derives the expected actual causes in most cases, including some where Woodwards and HP failed. In cases where actual P-cause fails, counterfactual reasoning is essential. Several examples can be tested at `http://adams.cs.kuleuven.be/idp/server.html?chapter=intro/11-AC`.

The aim to study actual causation in the context of causal processes is present in neuron diagrams approaches [23]. However, neuron diagrams do not represent individual mechanisms (similar to a structural equation) and do not distinguish between triggering and enabling conditions, and hence fall short for the sort of examples that motivated this paper. There exist other languages with a syntactic rule notation to express causal knowledge [28, 6, 5, 7]. However, it is not clear to us whether our view of causal mechanisms matches with the view of causal rules in some of these formalisms. The only causal reasoning study that accounts for

causal mechanisms, processes and worlds that we are aware of is the work on CP-logic [28]. CP-logic was used for various 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 there is no distinction between triggering and enabling conditions in CP-logic. The actual causation method for CP-logic proposed by Vennekens [27] and refined by Beckers and Vennekens [2] 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 [13, 12]. The relation with our approach is not obvious and we leave a further analysis of this for future work.

Several other topics for future research exist. One is to determine the complexity of key computational problems, such as computing different notions of actual causation. Useful extensions of the language include non-deterministic, probabilistic and cyclic causation, first order features (e.g., quantification), and dynamic mechanisms that initiate some property at one time and terminate it at another. This we plan to do following CP-logic, which supports several of these extensions. Another challenge is to develop a proof-theoretical account of the logic.

Acknowledgements

We thank Alexander Bochman, Sander Beckers, Jorge Fandinno, Mathieu Beirlaen, and anonymous reviewers for many discussions and valuable feedback.

References

1. Baumgartner, M.: A regularity theoretic approach to actual causation. Erkenn **78(Suppl) 1:85** (2013), https://doi.org/10.1007/s10670-013-9438-3
2. Beckers, S., Vennekens, J.: Counterfactual dependency and actual causation in CP-logic and structural models: A comparison. In: Proceedings of STAIRS. pp. 35–46 (2012)
3. Beckers, S., Vennekens, J.: A general framework for defining and extending actual causation using CP-logic. International Journal of Approximate Reasoning **77**, 105–126 (2016)
4. Beckers, S., Vennekens, J.: A principled approach to defining actual causation. Synthese **195**(2), 835–862 (2018). https://doi.org/10.1007/s11229-016-1247-1, https://doi.org/10.1007/s11229-016-1247-1
5. Bochman, A.: Actual causality in a logical setting. In: IJCAI (2018)
6. Bochman, A., Lifschitz, V.: Pearl's causality in a logical setting. In: Bonet, B., Koenig, S. (eds.) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 1446–1452. AAAI Press (2015), http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9686
7. Cabalar, P., Fandinno, J.: Enablers and inhibitors in causal justifications of logic programs. TPLP **17**(1), 49–74 (2017). https://doi.org/10.1017/S1471068416000107, https://doi.org/10.1017/S1471068416000107

8. Clark, K.L.: Negation as failure. In: Logic and Data Bases. pp. 293–322. Plenum Press (1978)
9. Fenton-Glynn, L.: A proposed probabilistic extension of the halpern and pearl definition of 'actual cause'. The British Journal for the Philosophy of Science (2015)
10. Gerstenberg, T., Goodman, N.D., Lagnado, D.A., Tenenbaum, J.B.: How, whether, why: Causal judgments as counterfactual contrasts. In: Proceedings of the 37th Annual Conference of the Cognitive Science Society. pp. 782–787 (2015)
11. Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Teng, C.M., Zhang, J.: Actual causation: a stone soup essay. Synthese **175**(2), 169–192 (2010)
12. Hall, N.: Structural equations and causation. Philosophical Studies **132**(1), 109–136 (2007)
13. Hall, N.: Two concepts of causation. In: Causation and Counterfactuals (2004)
14. Halpern, J.: Actual causality. MIT Press (2016)
15. Halpern, J., Pearl, J.: Causes and explanations: A structural-model approach. part i: Causes. The British Journal for the Philosophy of Science **56**, 843–87 (2005)
16. Halpern, J.Y.: Appropriate causal models and the stability of causation. Rew. Symb. Logic **9**(1), 76–102 (2016)
17. Hiddleston, E.: Causal powers. British Journal for the Philosophy of Science **56**(1), 27–59 (2005)
18. Hiddleston, E.: A causal theory of counterfactuals. Noûs **39**(4), 632–657 (2005)
19. Hitchcock, C.: The intransitivity of causation revealed in equations and graphs. Journal of Philosophy **98**, 273–299 (2001)
20. Hitchcock, C.: Prevention, preemption, and the principle of sufficient reason. Philosophical Review **116**(4), 495–532 (2007)
21. Hume, D.: A Treatise of Human Nature. John Noon (1739)
22. Lewis, D.: Causation. Journal of Philosophy **70**, 113–126 (1973)
23. Lewis, D.: Postscripts to 'causation'. In: Lewis, D. (ed.) Philosophical Papers Vol. Ii. Oxford University Press (1986)
24. McDermott, M.: Redundant causation. British Journal for Philosophy of Science **XLVI: 523-44** (1995)
25. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
26. Schaffer, J.: Causes need not be physically connected to their effects: The case for negative causation. In: Hitchcock, C.R. (ed.) Contemporary Debates in Philosophy of Science, pp. 197–216. Blackwell (2004)
27. Vennekens, J.: Actual causation in cp-logic. Theory and Practice of Logic Programming **11**, 647–662 (2011)
28. Vennekens, J., Denecker, M., Bruynooghe, M.: CP-logic: A language of causal probabilistic events and its relation to logic programming. TPLP **9**(3), 245–308 (2009)
29. Weslake, B.: A partial theory of actual causation. The British Journal for the Philosophy of Science (2015)
30. Woodward, J., Woodward, J., Press, O.U.: Making Things Happen: A Theory of Causal Explanation. Oxford scholarship online, Oxford University Press (2003), `https://books.google.be/books?id=LrAbrrj5te8C`