Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Actual Causality in a Logical Setting

Alexander Bochman
Computer Science Department, Holon Institute of Technology, Israel
bochmana@hit.ac.il

Abstract

We provide a definition of actual causation in the logical framework of the causal calculus, which is based on a causal version of the well-known NESS (or INUS) condition. We compare our definition with other, mainly counterfactual, approaches on standard examples. On the way, we explore general capabilities of the logical representation for structural equation models of causation and beyond.

1 Introduction

Studies of causation have revealed an important distinction between general (type-level) causation that deals with causal laws or law-like regularities, and actual (token) causation that involves singular causal claims *'c was a cause of e'*, where c and e are particular events. In this study we will use the causal calculus [Giunchiglia *et al.*, 2004; Bochman, 2003] as a logical formalism of general causation, and provide a formal description of actual causation in this framework.

Much recent work on actual causation is conducted within the structural equation framework [Pearl, 2000]. It has been shown in [Bochman and Lifschitz, 2015], however, that structural equation models are representable in the causal calculus, and we will make use of this representation for a logical translation of the examples of actual causation, described in this literature. We will not adopt, however, the dominant counterfactual approach to analyzing this notion. Instead, we will return to the traditional regularity approach, an approach "that has unjustly fallen into disfavor in some quarters" [Paul and Hall, 2013]. Our suggested definition of actual causation will be a particular instantiation of the INUS (or NESS) test, though elevated to causal language of the causal calculus.

The plan of the paper is as follows. First, we briefly describe the causal calculus and how structural equation models are representable in it. Then we provide a formal definition of actual causation in this setting. We will show, in particular, that the difference between general and actual causation could even be recast as a difference in their underlying logics. Then we proceed to some key examples and counterexamples of actual causation discussed in the literature, and show that the logical language of the causal calculus has representational capabilities that go beyond structural equation models.

2 General Causation in the Causal Calculus

The causal calculus was introduced in [McCain and Turner, 1997] as a nonmonotonic formalism for reasoning about action and change in AI. It forms a basis of the action description language $\mathcal{C}$ [Giunchiglia *et al.*, 2004]. A logical basis of the causal calculus was described in [Bochman, 2003], while [Bochman, 2004] studied its possible uses as a general-purpose nonmonotonic formalism.

In this study, we will use the causal calculus as a general logical formalism of causal reasoning. As such, it shares a common starting point with Pearl's approach to causality in that our knowledge can be stored in terms of cause-effect relationships. In the causal calculus, the latter are represented directly by causal rules of the form $A \Rightarrow B$ (*"A causes B"*), where A and B are classical propositions. Structural equation models are representable using such rules, so the approach can be viewed as a logical generalization of the latter.

Causal rules represent general (type-level) causal claims, so they correspond to such notions as nomic or causal sufficiency, causal laws and lawlike regularities. Just as the latter, causal rules are inherently *modal* notions.

A plausible way of interpreting causal rules consists in viewing them as representing (causal) *mechanisms* (cf. [Pearl, 2000]). This interpretation will play an important role in our approach to actual causation, though it will be based on a more fine-grained understanding of mechanisms than what is usually assumed in structural equation models.

Our basic language will be an ordinary propositional language with the classical connectives and constants $\{\wedge, \vee, \neg, \rightarrow, \mathbf{t}, \mathbf{f}\}$. $\models$ will stand for the classical entailment, while Th will denote the classical provability operator. We will often identify a propositional interpretation ('world') with the set of propositional formulas that hold in it.

In what follows, by a *causal theory* we will mean an arbitrary set of causal rules.

2.1 Nonmonotonic Semantics

A causal theory determines the set of situations (or worlds) that satisfy the rules of the theory. However, a distinctive feature of causal reasoning is that the relevant situations are determined not only by the rules that belong to the causal

theory, but also by what does *not* belong to it[1]. Accordingly, this principal semantic function is realized in the causal calculus by assigning a causal theory a particular *nonmonotonic* semantics: situations that satisfy a causal theory should also comply with Leibniz's Principle of Sufficient Reason - nothing happens without a sufficient reason, why it should be so.

For a causal theory Δ and a set u of propositions, let $\Delta(u)$ denote the set of propositions that are caused by u in Δ:

$$\Delta(u) = \{B \mid A \Rightarrow B \in \Delta, \text{ for some } A \in u\}$$

Definition 1.
- A consistent set u of propositions is an *exact model* of a causal theory Δ if $u = \mathrm{Th}(\Delta(u))$.
- A *general nonmonotonic semantics* of a causal theory is the set of all its exact models.
- A *causal nonmonotonic semantics* of a causal theory is the set of its exact models that are worlds.

An exact model is not only closed with respect to the causal rules; every proposition in it is also caused by other propositions that hold. The nonmonotonic semantics are indeed nonmonotonic, since adding new rules to a causal theory may lead to a nonmonotonic change of the semantics, and thereby to a nonmonotonic change in the derived information.

The causal nonmonotonic semantics is equivalent to the semantics introduced in [McCain and Turner, 1997].

2.2 Regular and Causal Inference

The causal calculus can be viewed as a two-layered construction. The nonmonotonic semantics form its top level. Its bottom level are logics of causal rules introduced in [Bochman, 2004]. A weakest such logic is a slight modification of the input-output logic from [Makinson and van der Torre, 2000]:

Definition 2. A *production inference relation* is a binary relation $\Rightarrow$ on the set of classical propositions satisfying the following conditions:

(Strengthening) If $A \vDash B$ and $B \Rightarrow C$, then $A \Rightarrow C$;

(Weakening) If $A \Rightarrow B$ and $B \vDash C$, then $A \Rightarrow C$;

(And) If $A \Rightarrow B$ and $A \Rightarrow C$, then $A \Rightarrow B \wedge C$;

(Truth and Falsity) $\mathbf{t} \Rightarrow \mathbf{t}$; $\mathbf{f} \Rightarrow \mathbf{f}$.

A characteristic property of production inference is that the reflexivity postulate $A \Rightarrow A$ does not hold for it.

Causal rules can be generalized to rules having arbitrary sets of propositions as premises using the familiar compactness recipe: for any set u of propositions,

$$u \Rightarrow A \;\equiv\; \bigwedge a \Rightarrow A, \text{ for some finite } a \subseteq u$$

A production inference relation is *regular* if it satisfies

(Cut) If $A \Rightarrow B$ and $A \wedge B \Rightarrow C$, then $A \Rightarrow C$.

Regular inference relations are already transitive. They will play an important role in describing actual causation.

A production inference relation is *basic* if it satisfies

(Or) If $A \Rightarrow C$ and $B \Rightarrow C$, then $A \vee B \Rightarrow C$.

[1] According to Pearl, causal assumptions are encoded in the missing links (that sanction, e.g., claims of zero covariance).

An important fact about basic production inference is that any causal rule is reducible to a set of *clausal* rules of the form $\bigwedge l_i \Rightarrow \bigvee l_j$, where l_i, l_j are classical literals.

Finally, a production inference relation is called *causal* if it is both basic and regular. Causal inference relations satisfy most of the usual postulates for classical entailment (except Reflexivity and Contraposition).

Nonmonotonic semantics indirectly determine their associated causal logics. More precisely, such a logic could be characterized as a maximal logic that preserves the nonmonotonic semantics under arbitrary expansions of a causal theory with additional causal rules. It has been shown in [Bochman, 2004] that causal inference is an adequate logic for the causal nonmonotonic semantics, whereas a weaker regular inference is adequate for the general nonmonotonic semantics.

2.3 Representing Structural Equations

In [Pearl, 2000, Chapter 7], a causal model was defined as a triple $M = \langle U, V, F\rangle$ where U is a set of *exogenous* variables, $V = \{V_i \mid i \leq n\}$ is a finite set of *endogenous* variables, and F is a set of functions such that each $f_i \in F$ is a mapping from $U \cup (V \backslash V_i)$ to V_i. F is represented as a set of equations

$$v_i = f_i(pa_i, u_i) \quad i = 1, \ldots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in $V \backslash \{V_i\}$ (parents) sufficient for representing f_i, and similarly for $U_i \subseteq U$. Each such equation is intended to represent a stable and autonomous physical mechanism, which means that it is conceivable to modify (or cancel) one such equation without changing the others.

Every instantiation $U = u$ of the exogenous variables determines a "causal world" of the causal model. Such worlds stand in one-to-one correspondence with the solutions to the above equations in the ordinary mathematical sense. However, structural equations also encode causal information in their very syntax by treating the variable on the left of $=$ as the effect and treating those on the right as causes. This causal reading plays a crucial role in determining the effect of external interventions and evaluation of counterfactuals.

For binary variables, Pearl's notion of a model can be formulated as follows (cf. [Bochman and Lifschitz, 2015]):

Definition 3. Assume that the set of propositional atoms is partitioned into a set of *exogenous* atoms and a finite set of *endogenous* atoms.

- A *Boolean structural equation* is an expression of the form $A = F$, where A is an endogenous atom and F is a propositional formula in which A does not appear.
- A *Boolean causal model* is a set of Boolean structural equations $A = F$, one for each endogenous atom A.

Definition 4. A *solution* (or a *causal world*) of a Boolean causal model M is any propositional interpretation satisfying the equivalences $A \leftrightarrow F$ for all equations $A = F$ in M.

[Bochman and Lifschitz, 2015] suggested the following translation of causal models into the causal calculus:

Definition 5. For any Boolean causal model M, Δ_M is the causal theory consisting of the rules

$$F \Rightarrow A \text{ and } \neg F \Rightarrow \neg A$$

for all equations $A = F$ in M and the rules

$$A \Rightarrow A \text{ and } \neg A \Rightarrow \neg A$$

for all exogenous atoms A of M.

This translation will be used in re-presenting structural equation models suggested for examples of actual causation.

3 Actual Causation Defined

Actual causation involves causal claims of the form *"C was a cause of E"*. In other words, it deals with *post factum* attribution of causal responsibility for actual outcome. Traditional regularity approach to this notion is exemplified by the well-known INUS condition[2]. A more adequate formulation has been suggested in [Wright, 1985]:

> The NESS test: a condition c was a cause of a consequence e if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of e.

Following [Lewis, 1973], however, the majority of authors have chosen a rival counterfactual approach to causation. A standard opinion in the literature has long been that regularity theories have unsurmountable difficulties. This opinion, however, is largely unjustified. To begin with, most of these difficulties can be met by adopting more stringent conditions on necessary and sufficient conditions (see [Baumgartner, 2013]). However, a more radical amendment has been suggested, e.g., in [Strevens, 2007], according to which the very notion of sufficiency (which has been assumed to be classical in the original regularity theory) should be given a causal interpretation. This view has been endorsed by the author of the NESS test himself in [Wright, 2011].

In the definition below, we will explicate the relevant notion of causal sufficiency in terms of causal inference[3].

3.1 Clausal Theories and Parsimony

Actual causation turns out to be highly sensitive to the syntactic form of causal rules. That is why we will require from the outset that the relevant causal theory should be a *clausal* theory, namely, it should consist only of causal rules of the form $l_1, \ldots, l_n \Rightarrow l$, where l and all l_i are literals.

In our approach to actual causation, *each* causal rule of a causal theory will be viewed as describing an autonomous causal mechanism[4]. This understanding presupposes, however, that the causal theory does not contain redundant causal rules that are subsumed by other rules:

Definition 6. A causal theory Δ will be called *parsimonious* (irredundant) if no causal rule from Δ is derivable from the rest of the rules in Δ by causal inference.

Suitable examples will be provided in what follows showing that the above constraints are essential for the correctness of actual causation claims in particular situations.

[2]Insufficient but Nonredundant part of an Unnecessary but Sufficient condition [Mackie, 1974].

[3]Such definitions have already been attempted in the framework of structural equation models - see, e.g., [Baldwin and Neufeld, 2004; Halpern, 2008].

[4]Cf. [Vennekens *et al.*, 2010] for a similar approach.

3.2 The Definition

On the account below, an actual causation claim presupposes a given causal theory Δ, and an actual world α that is a causal (exact) world with respect to Δ.

Definition 7. Let α be a causal world of a clausal causal theory Δ. A causal rule $l_1, \ldots, l_n \Rightarrow l$ will be called *active in* α if $\{l_1, \ldots, l_n\} \subseteq \alpha$. The *actual sub-theory* of Δ wrt α is the set Δ_α of all causal rules from Δ that are active in α.

Since causal worlds of Δ are closed with respect to the rules of Δ, the heads of all causal rules that are active in α will also hold in α. Note that any causal world is uniquely determined by the causal rules that are active in it.

In what follows, $\Rightarrow_\alpha$ will denote the least causal inference relation that includes Δ_α.

Definition 8 (actual cause). Let α be a causal world of a parsimonious clausal causal theory Δ. A literal $l_0 \in \alpha$ will be said to be an *actual cause* of a literal l in α wrt Δ if and only if there exists a set of literals $L \subseteq \alpha$ such that

1. $l_0, L \Rightarrow_\alpha l$;
2. $L \not\Rightarrow_\alpha l$.

The above definition provides a direct formalization of the NESS test by defining sufficiency as causal inference in the actual sub-theory of the source causal theory.

There is a lot of similarity between the actual causal sub-theory and the notion of a causal beam in [Pearl, 2000]. Our definition has also much in common with the approach of [Beckers and Vennekens, 2018]. In particular, their notion of production can be viewed as a counterpart of our logical notion of causal inference $\Rightarrow_\alpha$ in the actual sub-theory.

On the suggested account, general and actual causation are conceptually different. Namely, general causation is a purely *logical* notion that is described by an appropriate formalism of causal inference. In contrast, actual causation is already an explicitly *nonmonotonic* notion, since it depends on the absence (non-provability) of certain causal rules. Unlike general causation, claims of actual causation can be overridden with an addition of new causal rules to a causal theory.

Yet another salient feature of the above definition is its high sensitivity to the syntactic form of causal rules[5]. For instance, suppose that disjunctions of literals are allowed in the bodies of causal rules, and assume that l_0 is an actual cause of l by the above definition: for some set L of literals,

$$l_0, L \Rightarrow_\alpha l \quad \text{and} \quad L \not\Rightarrow_\alpha l$$

Now let r be an arbitrary literal from α. Then we have

$$r, (\neg r \vee l_0), L \Rightarrow_\alpha l \quad \text{and} \quad (\neg r \vee l_0), L \not\Rightarrow_\alpha l$$

by the logical properties of $\Rightarrow_\alpha$. So, if we could add $\neg r \vee l_0$ to the 'witness' set L, we would obtain that r is an actual cause of l, which is absurd.

It should be kept in mind, however, that this syntax sensitivity is not specific to our definition, or even to regularity accounts in general. Rather, it might be a feature of the notion of actual causation itself. Starting with the stipulation in

[5]As a matter of fact, our definition 'inherited' its sensitivity to the syntax from the original INUS condition (see [Strevens, 2007]).

[Lewis, 1973] that causal relata are primitive 'events', practically all counterfactual accounts of causation impose similar restrictions. There are severe differences in opinions even about whether 'negations' (absences and omissions) of events can be causal relata, let alone 'disjunctive' events.

3.3 Actual Causation and Regular Inference

The following key result will show that causal inference with respect to the actual sub-theory $\Rightarrow_\alpha$ can be replaced with an unconstrained *regular* inference. This will recast part of the difference between general and actual causation as a difference in their underlying logics.

Let $\Rightarrow^r_\Delta$ denote the least regular inference relation containing a causal theory Δ. Then we have

Theorem 1. *Let α be a causal world of a clausal causal theory Δ. Then for any $L \subseteq \alpha$ and any literal l,*

$$L \Rightarrow_\alpha l \text{ iff } L \Rightarrow^r_\Delta l.$$

As a consequence, we obtain the following equivalent characterization of actual causation:

Corollary 2. *Let α be a causal world of a clausal causal theory Δ. Then $l_0 \in \alpha$ is an actual cause of l in α wrt Δ if and only if there exists a set of literals $L \subseteq \alpha$ such that*

1. $l_0, L \Rightarrow^r_\Delta l$;
2. $L \not\Rightarrow^r_\Delta l$.

The above description makes our definition of actual causation a straightforward formalization of the NESS test with regular inference as a logical explication of (causal) sufficiency. As a consequence, regular inference is adequate for reasoning about actual causation:

Corollary 3. *Regularly equivalent clausal causal theories support the same claims of actual causation.*

In the next section we will test our definition on a number of standard examples in the literature.

4 Examples and Counterexamples

The examples we are going to discuss occupy a prominent place in the literature, mainly because each of them constitutes a counterexample for some past approach to actual causation. For the majority of these examples, there are established representations in structural equation models (see, e.g., [Halpern, 2016]) , and we will use them as a basis of our logical characterization. This will not mean, of course, that our suggested definition will always produce the same answers, though there will indeed be a large area of agreement.

To begin with, the role of the restriction to the actual subtheory can be illustrated on the following example.

Example 1 (**Loader** [Hopkins and Pearl, 2003]). A firing squad consists of shooters B and C. It is A's job to load B's gun, C loads and fires his own gun. On a given day, A loads B's gun. When the time comes, only C shoots the prisoner.

The initial definition in [Halpern and Pearl, 2001] wrongly made A an actual cause of D.

The structural equation for this example is as follows:

$$D = (A \wedge B) \vee C$$

It corresponds to the following clausal causal theory:

$$A, B \Rightarrow D \quad C \Rightarrow D$$
$$\neg A, \neg C \Rightarrow \neg D \quad \neg B, \neg C \Rightarrow \neg D$$

For the actual causal world $\{A, \neg B, C, D\}$, the associated actual causal theory is just $\{C \Rightarrow D\}$. Thus, C is clearly an actual cause of D, but A is not.

Examples of overdetermination and preemption present *prima facie* problems for counterfactual theories of causation, because in such cases there is no direct counterfactual dependence between the effect and its cause.

Symmetric overdetermination.

Example 2 (**Window**). Billy (B) and Suzy (S) both throw rocks at a window. The rocks strike the window at exactly the same time. The window breaks (W).

The structural equation of this story is just $W = B \vee S$, so the corresponding causal theory is as follows

$$B \Rightarrow W \quad S \Rightarrow W \quad \neg B, \neg S \Rightarrow \neg W.$$

The actual causal world is $\alpha = \{S, B, W\}$, and only the first two rules are active in it. We have both $B \Rightarrow_\alpha W$ and $S \Rightarrow_\alpha W$, though $\mathbf{t} \not\Rightarrow_\alpha W$, and therefore both S and B are actual causes of W in α.

Early preemption.

Example 3 (**Backup** [Hitchcock, 2007]). Assassin poisons Victim's coffee (A). Victim drinks it and dies (D). If Assassin hadn't poisoned the coffee, Backup would have (B), and Victim would have died anyway.

Backup is represented using the equations:

$$B = \neg A, \quad D = A \vee B$$

The latter correspond to the following causal theory Δ:

$$\neg A \Rightarrow B \qquad A \Rightarrow \neg B$$
$$A \Rightarrow D \quad B \Rightarrow D \quad \neg A, \neg B \Rightarrow \neg D$$

$\neg A \Rightarrow B$ and $B \Rightarrow D$ imply $\neg A \Rightarrow D$. Taken together with $A \Rightarrow D$, the latter implies $\mathbf{t} \Rightarrow D$ by Or. Thus, on the level of general causation, D is 'causally inevitable'.

For the actual world α=$\{A, \neg B, D\}$, the actual causal theory is $\{A \Rightarrow \neg B, A \Rightarrow D\}$. We have $A \Rightarrow_\alpha D$, though $\mathbf{t} \not\Rightarrow_\alpha D$, and therefore A is an actual cause of D in α.

A distinctive feature of our account is that in the situation where Assassin does not poison the coffee (that is $\alpha = \{\neg A, B, D\}$), $\neg A$ is an actual cause of D (as well as B), since in this case the actual causal theory is

$$\neg A \Rightarrow B \quad B \Rightarrow D$$

Surprisingly, the last, modified definition from [Halpern, 2016] does not support this claim. Even more surprisingly, just as in the original counterfactual account of [Lewis, 1973], the claim is restored if we add some intermediate event (say 'Victim drinks the poisoned coffee') on the causal path from A to D. We agree here with [Hitchcock, 2001] that we should be particularly troubled that we judge there to be a causal relationship on the basis of finding an intermediate event that is not made salient in the presentation of the example.

Remark. In the sequel to this study (forthcoming), it will be shown that the above configuration practically exhausts the area of 'positive' disagreement between our definition and that of Halpern; otherwise, it can be proved that if C is an actual cause of E on our account, it will be a(t least part of a) cause of E on Halpern's modified definition. Due to the results stated in [Halpern, 2016], this will also imply that our definition is more 'conservative' than practically any other counterfactual account. Still, there are cases where what is an actual cause on all counterfactual accounts will not be an actual cause on our definition - see the *Shock* example below.

Late preemption.

Example 4 (**Bottle**). Suzy (ST) and Billy (BT) both throw rocks at a bottle, Suzy's rock arrives first and hits the bottle (SH), the bottle shatters (BS), Billy's arrives second and so does not hit the bottle (BH). Both throws are accurate, Billy's would have shattered the bottle if Suzy's had not.

The following structural equation model has been suggested in [Halpern and Pearl, 2001]:

$$SH = ST, \quad BH = BT \wedge \neg SH, \quad BS = BH \vee SH$$

The corresponding causal theory is:

$$ST \Rightarrow SH \quad BT, \neg SH \Rightarrow BH \quad BH \Rightarrow BS \quad SH \Rightarrow BS$$
$$\neg ST \Rightarrow \neg SH \quad \neg BH, \neg SH \Rightarrow \neg BS$$
$$\neg BT \Rightarrow \neg BH \quad SH \Rightarrow \neg BH$$

For general causation, we have again an overdetermination, namely not only $ST \Rightarrow BS$, but also $BT \Rightarrow BS$! However, for the actual world $\alpha = \{ST, BT, SH, \neg BH, BS\}$, the corresponding actual causal theory is

$$ST \Rightarrow SH \quad SH \Rightarrow BS \quad SH \Rightarrow \neg BH$$

We have $ST \Rightarrow BS$ by transitivity, so both SH and ST are actual causes of BS in α, as expected. It is clear also that BT cannot be an actual cause of D in this world.

The above description of bottle shattering involves auxiliary variables SH and BH whose only role consists in enabling a counterfactual description of the difference between preempting and preempted cause. A simpler description of the situation could as well be as follows:

Example 5 (**Simplified Bottle**). Suzy (ST) and Billy (BT) both throw rocks at a bottle, but Suzy's rock arrives first and shatters the bottle (BS). Both throws are accurate: Billy's would have shattered the bottle if Suzy's had not.

Though it seems there is no apt structural model just on the salient variables of this simplified description, there is a simple causal theory that describes them:

$$ST \Rightarrow BS \quad BT, \neg ST \Rightarrow BS \quad \neg ST, \neg BT \Rightarrow \neg BS$$

For the variables involved, this theory support the same claims of actual causation. The adequacy of this causal theory in describing preemption is based, however, on the fact that the set of causal rules $\{ST \Rightarrow BS,\ \neg ST, BT \Rightarrow BS\}$ is logically distinct from $\{ST \Rightarrow BS,\ BT \Rightarrow BS\}$ for regular inference, though they are equivalent with respect to causal inference.

'Structural' non-equivalence. The existence of regularly different logical representations for causally equivalent descriptions can also be exploited for resolving the problem of 'structural equivalents' in structural equation models.

Example 6 (**Bogus Prevention** [Hiddleston, 2005]). Assassin refrains from putting poison in Victim's coffee ($\neg A$). Bodyguard puts antidote in the coffee (B). Victim drinks the coffee and survives ($\neg D$).

A seemingly appropriate equation $D = A \wedge \neg B$ makes this case 'isomorphic' to Window (symmetric overdetermination), with a counter-intuitive conclusion that both $\neg A$ *and* B are actual causes of $\neg D$.

The intuitive asymmetry of these potential causes could be captured by using auxiliary variables (e.g., poison neutralization)[6]. In our logical framework, however, the same effect can be achieved using the following causal theory:

$$A, \neg B \Rightarrow D \quad \neg A \Rightarrow \neg D \quad A, B \Rightarrow \neg D$$

The last rule provides a formal description of poison neutralization. For the actual world $\alpha = \{\neg A, B, \neg D\}$, the corresponding actual causal theory is just $\{\neg A \Rightarrow \neg D\}$, and consequently only $\neg A$ is an actual cause of $\neg D$.

Switches. Preemption examples invariably involve a pattern where some variable acts as a switch between two mechanisms or processes, both leading to the same result. Accordingly, this variable does not influence the result on the general causal level, though its actual instantiations are actual causes of this result in each particular situation. We will discuss below a couple of more complex examples of this kind.

Example 7 (**Push** [McDermott, 1995]). I push (P) Jones in front of a truck (T), which hits (H) and kills him (D); if I had not done so, a bus (B) would have hit and killed him.

Below is a corresponding structural model:

$$H = (P \wedge T) \vee (\neg P \wedge B), \quad D = H$$

The HP definition from [Halpern and Pearl, 2005] yielded P and T as causes of D, as we would expect. But, unfortunately, it also yielded B as a cause of D.

The 'positive' part of the corresponding causal theory is[7]

$$P, T \Rightarrow H \quad \neg P, B \Rightarrow H \quad H \Rightarrow D$$

The actual world is $\{P, B, T\}$, so the actual theory is

$$P, T \Rightarrow H \quad H \Rightarrow D$$

Thus, P and T are actual causes of D, but B is not.

The above example can also be used to illustrate the necessity of a parsimonious (non-redundant) representation of causal mechanisms. Note that the first two rules imply the following rule by causal inference:

$$T, B \Rightarrow H.$$

Actually, this is an immediate consequence of a purely logical fact that $(P \wedge T) \vee (\neg P \wedge B)$ is equivalent to

$$(P \wedge T) \vee (\neg P \wedge B) \vee (T \wedge B)$$

[6]This solution has been suggested in [Blanchard and Schaffer, 2017] and mentioned in [Halpern and Hitchcock, 2015].

[7]The rest of the causal theory is irrelevant for this example.

in classical logic. However, if we would add the above causal rule to the source causal theory, it would appear also in the actual causal theory, and we would obtain that also B is an actual cause of D, contrary to our intuitions.

In the example below what is a cause on any counterfactual approach will not be an actual cause on our definition[8].

Example 8 (**Inevitable Shock** [McDermott, 1995; Weslake, 2015]). Two switches are wired to an electrode. The switches are controlled by A and B respectively, and the electrode is attached to C. A flips her switch (A), which *forces* B to flip her switch (B) (B has no other option). The electrode is activated and shocks C (C) iff both switches are in the same position.

The corresponding causal theory is

$$A \Rightarrow B \quad A, B \Rightarrow C \quad \neg A, \neg B \Rightarrow C$$
$$\neg A \Rightarrow \neg B \quad \neg A, B \Rightarrow \neg C \quad A, \neg B \Rightarrow \neg C$$

Again, $\mathbf{t} \Rightarrow C$ follows from the above theory by causal inference, so the shock is inevitable. Still, the actual world is $\{A, B, C\}$, so the actual causal sub-theory is

$$A \Rightarrow B \quad A, B \Rightarrow C$$

We obtain $A \Rightarrow C$ by Cut, so B cannot be an actual cause of C (since any set of literals that causes C will include A). This makes A the *only* actual cause of C. In fact, the source causal theory is *regularly* equivalent to the following one:

$$A \Rightarrow B \quad A \Rightarrow C \quad \neg A \Rightarrow \neg B \quad \neg A \Rightarrow C$$

In the above theory, B and C are just joint effects of the common cause A. Note, however, that B is a but-for cause of C in the original causal theory (due to the last rule, $A, \neg B \Rightarrow \neg C$), so it is an actual cause of the latter on any counterfactual account. The regularity account provides here more discriminate answers about actual causality than the counterfactual approach.

(In)transitivity of causation. One of the most discussed features of causation is transitivity, and the suggested theory allows to explain at least some of the deliberations arising about this controversial property. In our theory, general causal inference is transitive, while actual causation is not. The following example from [Ehring, 1987] forms perhaps the simplest counterexample to transitivity on our definition:

Example 9 (**Purple flame**). Jones puts potassium salts (P) into a hot fire (F). Because potassium compounds produce a purple flame when heated, the flame changes to a purple colour (PF), though everything else remains the same. Both flames ignite some flammable material (I).

$$P, F \Rightarrow PF \quad F \Rightarrow I \quad PF \Rightarrow I$$
$$\neg P \Rightarrow \neg PF \quad \neg F \Rightarrow \neg PF \quad \neg F, \neg PF \Rightarrow \neg I$$

If the actual world α is $\{P, PF, F, I\}$, then the actual causal sub-theory is

$$P, F \Rightarrow PF \quad F \Rightarrow I \quad PF \Rightarrow I,$$

and hence P is an actual cause of PF, and PF is an actual cause of I. However, P is not an actual cause of I, since it is not a necessary part of any sufficient condition for I.

[8]We have slightly changed the story to make it more in accord with the equations (specifically, $B := A$ used in [Weslake, 2015]).

4.1 Summary and Prospects

Causation is a notoriously difficult and complex notion. In our logical approach, part of its complexity is reflected in the fact that the causal calculus is not a plain logical system with stipulated axioms, but an essentially nonmonotonic formalism in which logics and nonmonotonic semantics are tightly intertwined. As we have seen, this representational complexity is even higher for actual causation. Still, our suggested definition has produced reasonably simple algorithms for checking the relevant causal claims.

There are at least two major representational issues that have been left outside the scope of this study. The first concerns the use of multi-valued (non-binary) variables that have shown their usefulness in the framework of structural equation models. The causal calculus has all the necessary tools for dealing with such variables, which creates ample opportunities for a formal representation of the relevant cases of actual causation (such as trumping preemption - see, e.g., [Weslake, 2015]).

The second, larger issue concerns the role and use of normality and defaults in reasoning about actual causation. Again, the causal calculus already has the relevant tools for a formal representation of these concepts (see [Giunchiglia *et al.*, 2004]), so it seems to suggest a promising approach for the study of the latter within a single logical framework.

Among more specific aims, an important objective of the study was to show that, once placed on proper logical grounds, traditional regularity approach to causation provides not only a natural, but also a viable definition of actual causation. In some sense, the viability of this approach lends an additional support to the counter-slogan from [Pearl, 2000]: "Causation without manipulation? You bet!"

Of course, our approach does not 'cancel' the counterfactual accounts, it only poses anew the old philosophical questions about the relation between (causal) laws and counterfactuals, questions that could be traced back to the famous double definition of causation by David Hume. Still, we can't help to agree with [Weslake, 2015] that there is a nice irony in the fact that most plausible counterfactual theories of causation turn out to draw heavily from the resources of the regularity theories they was initially motivated by rejecting.

We are intending to provide a more systematic comparison between our approach and counterfactual accounts in the sequel of this study [forthcoming], which, in turn, could be viewed as part of a larger, and independently important, logical study of causal counterfactuals in the framework of the causal calculus.

Actual causation, the primary subject of this study, is only part of the bigger picture of causality. In this respect, a larger aim of this study consisted in demonstrating that the causal calculus provides a unifying logical framework for causal reasoning. We believe that the study lends one more piece of justification for the use of logical tools and representations in the study of causation.

Acknowledgment. I am grateful to Vladimir Lifschitz for his comments on an earlier version of this paper.

References

[Baldwin and Neufeld, 2004] R. A. Baldwin and E. Neufeld. The structural model interpretation of the NESS test. In *Advances in Artificial Intelligence*, volume 3060 of *Lecture Notes in Computer Science*, pages 297–307. Springer, 2004.

[Baumgartner, 2013] M. Baumgartner. A regularity theoretic approach to actual causation. *Erkenntnis*, 78:85–109, 2013.

[Beckers and Vennekens, 2018] S. Beckers and J. Vennekens. A principled approach to defining actual causation. *Synthese*, 195:835–862, 2018.

[Blanchard and Schaffer, 2017] T. Blanchard and J. Schaffer. Cause without default. In H. Beebee, C. Hitchcock, and H. Price, editors, *Making a Difference*, pages 175–214. Oxford University Press, 2017.

[Bochman and Lifschitz, 2015] A. Bochman and V. Lifschitz. Pearl's causality in a logical setting. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 1446–1452. AAAI Press, 2015.

[Bochman, 2003] A. Bochman. A logic for causal reasoning. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 141–146, Acapulco, 2003. Morgan Kaufmann.

[Bochman, 2004] A. Bochman. A causal approach to nonmonotonic reasoning. *Artificial Intelligence*, 160:105–143, 2004.

[Ehring, 1987] D. Ehring. Causal relata. *Synthese*, 73:319–328, 1987.

[Giunchiglia *et al.*, 2004] E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, and H. Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153:49–104, 2004.

[Halpern and Hitchcock, 2015] J. Halpern and C. Hitchcock. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66(2):413–457, 2015.

[Halpern and Pearl, 2001] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach—part I: Causes. In *Proc. Seventh Conf. On Uncertainty in Artificial Intelligence (UAI'01)*, pages 194–202, San Francisco, CA, 2001. Morgan Kaufmann.

[Halpern and Pearl, 2005] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part I: Causes. *British Journal for Philosophy of Science*, 56(4):843–887, 2005.

[Halpern, 2008] J. Y. Halpern. Defaults and normality in causal structures. In *Principles of Knowledge Representation and Reasoning: Proc. Eleventh International Conference (KR'08)*, pages 198–208, 2008.

[Halpern, 2016] J. Halpern. *Actual Causality*. MIT Press, Cambridge, MA, 2016.

[Hiddleston, 2005] E. Hiddleston. Causal powers. *British Journal for Philosophy of Science*, 56:27–59, 2005.

[Hitchcock, 2001] C. Hitchcock. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, XCVIII(6):273–299, 2001.

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

[Hopkins and Pearl, 2003] M. Hopkins and J. Pearl. Clarifying the usage of structural models for commonsense causal reasoning. In *Proc. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2003.

[Lewis, 1973] D. Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973.

[Mackie, 1974] J. L. Mackie. *The Cement of the Universe. A Study of Causation*. Clarendon Press, Oxford, 1974.

[Makinson and van der Torre, 2000] D. Makinson and L. van der Torre. Input/Output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[McCain and Turner, 1997] N. McCain and H. Turner. Causal theories of action and change. In *Proceedings AAAI-97*, pages 460–465, 1997.

[McDermott, 1995] M. McDermott. Redundant causation. *British Journal for the Philosophy of Science*, 46(4):523–544, 1995.

[Paul and Hall, 2013] L. A. Paul and N. Hall. *Causation: A User's Guide*. Oxford University Press, 2013.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoninig and Inference*. Cambridge UP, 2000. 2nd ed., 2009.

[Strevens, 2007] M. Strevens. Mackie remixed. In J. K. Campbell, M. O'Rourke, and H. S. Silverstein, editors, *Causation and Explanation*. MIT Press, Cambridge MA, 2007.

[Vennekens *et al.*, 2010] J. Vennekens, M. Bruynooghe, and M. Denecker. Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In T. Janhunen and I. Niemelä, editors, *Logics in Artificial Intelligence: 12th European Conference, JELIA 2010, Helsinki, Finland, September 13-15, 2010. Proceedings*, pages 313–325, Berlin, Heidelberg, 2010. Springer.

[Weslake, 2015] B. Weslake. A partial theory of actual causation. *British Journal for the Philosophy of Science*, 2015. To appear.

[Wright, 1985] R. W. Wright. Causation in tort law. *California Law Review*, 1735:1788–91, 1985.

[Wright, 2011] R. W. Wright. The NESS account of natural causation: A response to criticisms. In R. Goldberg, editor, *Perspectives on Causation*, chapter 14. Hart Publishing, 2011. Available at SSRN: https://ssrn.com/abstract=1918405.