Situation Calculus Semantics for Actual Causality

Vitaliy Batusov
York University
Toronto, Canada
vbatusov@cse.yorku.ca

Mikhail Soutchanski
Ryerson University
Toronto, Canada
mes@scs.ryerson.ca

Abstract

The state-of-the-art definitions of actual cause by Pearl and Halpern suffer from the modest expressivity of causal models. We develop a new definition of actual cause in the context of situation calculus (SC) basic action theories. As a result, we avoid the paradoxes that arise in causal models and can identify complex actual causes of conditions expressed in first-order logic. We provide a formal translation from causal models to SC and establish a relationship between the definitions. Using examples, we show that long-standing disagreements between alternative definitions of actual causality can be mitigated by faithful modelling.

1 Introduction

Actual causality, also known as token-level causality, is concerned with finding in a given scenario a singular event that caused another event. This is in contrast to type-level causality which is concerned with universal causal mechanisms governing the world. The leading line of computational inquiry into actual causality was pioneered by [Pearl, 1998; 2000] and continued by [Halpern and Pearl, 2005; Halpern, 2000; Eiter and Lukasiewicz, 2002; Hopkins, 2005; Halpern, 2015; 2016] and in other publications. We call it the *HP approach*. It is based on the concept of structural equations [Simon, 1977] and implemented in the framework of causal models. The HP approach follows the Humean counterfactual definition of causation, which posits that saying "an event A caused an outcome B" is the same as saying "if A had not been, then B never had existed". This definition is well-known to suffer from the problem of *preemption*: it could be the case that in the absence of event A, B would still have occurred due to another event, which in the original scenario was preempted by A. HP address this by performing counterfactual analysis only under carefully selected contingencies which suspend some subset of the model's mechanisms. Selecting proper contingencies proved to be a challenging task; as mentioned in [Halpern, 2016] on p.27, "The jury is still out on what the 'right' definition of causality is".

The HP approach is prone to producing results that cannot be reconciled with intuitive understanding due to the limited expressiveness of causal models [Hopkins, 2005; Hopkins and Pearl, 2007]. The ontological commitments of structural causal models resemble propositional logic, they have no objects, no relationships, no time, no support for quantified causal queries. As a remedy, [Hopkins, 2005; Hopkins and Pearl, 2007] leverage the expressive power of first-order logic and the robustness of the situation calculus (SC) [Reiter, 2001]. To formulate counterfactuals within SC, they allow arbitrary modifications in a sequence of actions, e.g. removing actions that serve as preconditions for subsequent actions. They do not define actual causality.

Given that theories of actual causality based on structural equations share the same ailments [Menzies, 2014; Glymour *et al.*, 2010], it seems natural to explore actual causality from a different perspective. We do this in the language of SC under the classical Tarskian semantics, where the notion of a cause naturally aligns with the notion of an action, and the effect can be specified by a FOL formula with quantifiers over object variables. In contrast to HP whose analysis is based on observing the end results of interventions, we do so by analyzing the dynamics which lead to the end results. Our developments are based on a small set of plausible intuitions.

The next section briefly summarizes SC. Section 3 motivates our approach and supplies a running example. Section 4 characterizes causes which *achieve* an effect. Section 5 explores *maintenance* causes—actions which protect existing conditions from being lost. In Section 6, we combine achievement and maintenance causes into an all-encompassing notion of actual cause. In Section 7, we outline the HP approach and, in Section 8, formally connect it to ours. Finally, we briefly compare our definition to others using examples and discuss related work.

2 Situation Calculus

In the situation calculus [McCarthy and Hayes, 1969; Reiter, 2001], the constant S_0 denotes the initial situation that represents an empty list of actions, while the complex situation term $do([\alpha_1, ..., \alpha_n], S_0)$ represents the situation that results from executing actions α_1, ..., α_n consecutively so that α_1 is executed in S_0, and α_n is executed last. If none of the action terms α_i have variables, then we call this situation term an (actual) *narrative*. An action term α_i may occur in the narrative more than once at different positions. The set of all situations can be visualized as a tree with a partial-order relation $s_1 \sqsubset s_2$ on situations s_1, s_2, and $s_1 \sqsubseteq s_2$ abbreviates $s_1 \sqsubset s_2 \lor s_1 = s_2$. It is characterized by the foundational domain-independent axioms (Σ) included in a basic action theory (BAT) $\mathcal{D}$ that also includes axioms $\mathcal{D}_{S_0}$ describing the initial situation, and action precondition axioms $\mathcal{D}_{ap}$ using the predicate $Poss(a, s)$ to say when an action a is possible in s. For each action function there is one precondition axiom $Poss(A(\bar{x}), s) \leftrightarrow \Pi_A(\bar{x}, s)$, where as usual, all free variables are implicitly $\forall$-quantified, and $\Pi(\bar{x}, s)$ is a formula *uniform* in s, meaning that it has no occurrences of $Poss, \sqsubset$, no other situation terms, no quantifiers over situations. For each fluent F, $\mathcal{D}$ includes a successor state axiom (SSA)

$$F(\bar{x}, do(a, s)) \leftrightarrow \psi^+(\bar{x}, a, s) \vee F(\bar{x}, s) \wedge \neg\psi^-(\bar{x}, a, s)),$$

where the fluent predicate $F(\bar{x}, s)$ represents a situation-dependent relation over tuple of objects $\bar{x}$, uniform formulas $\psi^+(\bar{x}, a, s)$ and $\psi^-(\bar{x}, a, s)$ specify action terms that under certain application-dependent conditions have a positive effect (make F true), or a negative effect on fluent F (make it false), respectively. The SSAs are derived under the causal completeness assumption [Reiter, 1991] that all effects of actions on fluents are explicitly represented. There are a number of auxiliary axioms, such as unique name axioms, that are included in $\mathcal{D}$. The common abbreviation $executable(s)$ means that each action mentioned in the situation term s was possible in the situation in which it was executed. The basic computational task, called the *projection problem*, is the task of establishing whether a BAT entails a sentence $\phi(\sigma)$ for an executable ground situation σ, where $\phi(s)$ is a formula uniform in s. This problem can be solved using the one-step regression operator ρ. The expression $\rho[\varphi, \alpha]$ denotes the formula obtained from φ by replacing each fluent atom F in φ with the right-hand side of the SSA for F where the action variable a is instantiated with the ground action α, and then simplified using unique name axioms for actions and constants. Similarly to the theorem about multi-step regression $\mathcal{R}$ presented in [Reiter, 1991], one can prove that given a BAT $\mathcal{D}$, a formula $\varphi(s)$ uniform in s, and a ground action term α, we have that $\mathcal{D} \models \forall s.\ \varphi(do(\alpha, s)) \leftrightarrow \rho[\varphi(s), \alpha]$.

3 Motivation

We propose to axiomatize a dynamic world using a SC theory and derive actual causality from first principles. Specifically, to represent a "scenario", we consider a BAT $\mathcal{D}$ and a narrative describing the actions or events which transpired in the world characterized by $\mathcal{D}$. We do not formally distinguish between agent actions and nature's events. The narrative is specified by an executable ground situation term σ called the "actual situation". An effect for which we seek to identify causes is given by a formula $\varphi(s)$ uniform in situation s. Since actions are the sole source of change in a BAT, we identify the set of potential causes of an effect φ with the set of all ground action terms occurring in σ.

Example 1. A *D flip-flop* is a digital circuit capable of storing one bit of information. A basic D flip-flop has two Boolean inputs, D and $enable$, and one output, Q. Each input and output signal can be either at the low level (modelled as *false*), or at the high level (modelled as *true*). If an input $enable$ is high, every time the $clock$ "ticks", the output Q assumes the value of the main input D and maintains it until the next tick. When the signal $enable$ is low, the flip-flop preserves the value of Q regardless of D and ticks.

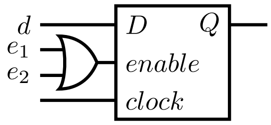

Figure 1: A D flip-flop with a disjunctive enable ($clock$ input is shown but not modeled)

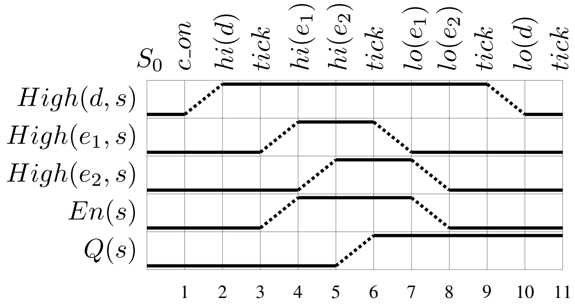

Figure 2: Evolution of fluent values throughout σ

Consider the circuit in Figure 1. It consists of a D flip-flop, shown as a box, whose $enable$ input is controlled by an OR-gate such that at least one of the signals e_1, e_2 needs to be high in order for the flip-flop to produce the output $Q = D$.

Let d, e_1, and e_2 be constants that represent the input signals. Let the action functions be $hi(x)$, $lo(x)$, $tick$, and c_on, where the first two actions set signal x to high or to low voltage level, respectively, $tick$ represents the action of the clock, and c_on turns the clock on, making $tick$ possible. The fluent $ClockOn(s)$ represents the state of the clock, $High(x, s)$ represents the logical value of signal x, $En(s)$ represents the output of the OR-gate, and $Q(s)$ is the output of the flip-flop.

Let the narrative σ be $do([c\_on, hi(d), tick, hi(e_1), hi(e_2),$ $tick, lo(e_1), lo(e_2), tick, lo(d), tick], S_0)$, and let the effect of interest be $Q(s)$. In the initial situation, we have that $\forall x(\neg High(x, S_0)), \neg Q(S_0), \neg En(S_0)$. The following BAT models the operation of the circuit.

$$\begin{aligned}
&Poss(tick, s) \leftrightarrow ClockOn(s), \quad Poss(c\_on, s),\\
&Poss(hi(x), s) \leftrightarrow \neg High(x, s),\\
&Poss(lo(x), s) \leftrightarrow High(x, s),
\end{aligned}$$

$$\begin{aligned}
&ClockOn(do(a, s)) \leftrightarrow a = c\_on \vee ClockOn(s),\\
&High(x, do(a, s)) \leftrightarrow a = hi(x) \vee High(x, s) \wedge a \neq lo(x),\\
&En(do(a, s)) \leftrightarrow a = hi(e_1) \vee a = hi(e_2) \vee\\
&\qquad En(s) \wedge \neg[a = lo(e_1) \wedge \neg High(e_2, s)] \wedge\\
&\qquad\quad \neg[a = lo(e_2) \wedge \neg High(e_1, s)],\\
&Q(do(a, s)) \leftrightarrow [a = tick \wedge En(s) \wedge High(d, s)] \vee\\
&\qquad Q(s) \wedge \neg[a = tick \wedge En(s) \wedge \neg High(d, s)].
\end{aligned}$$

Figure 2 graphically shows the truth values, relative to $\mathcal{D}$, of the key ground fluents in situation S_0 and after each subsequent action in σ. Observe that all fluents are initially false, shown as the thick lower edges, the #1 action c_on makes subsequent $tick$ actions (#3, #6, #9, #11) possible, the actions $hi(d)$, $hi(e_1)$, $hi(e_2)$, $lo(e_1)$ change the voltage levels of the corresponding signals, $hi(e_1)$ also changes the state of

$En(s)$, the second occurrence of $tick$ (#6) makes the output $Q(s)$ true, but other occurrences of $tick$ are inconsequential.

It is obvious that the 6-th action, $tick$, is a cause of $Q(s)$ in σ, having acted as the proverbial last straw that broke the camel's back, but so are the actions $hi(d)$ and $hi(e_1)$, having created the right circumstances for the back-breaking. Action #6 would accomplish nothing had the flip-flop not been enabled and the input bit set to high. The task before us is to introduce general formal criteria for identifying such actions.

We axiomatically recognize two kinds of causal roles which events may assume. *Achievement causes* are the events which realize—in whole or in part—either the condition of interest or the preconditions of other achievement causes. *Maintenance causes* are the events which prevent other events from falsifying the condition of interest. We use the generic term *actual cause* to refer to an event which contributes to the effect of interest via a combination of these causal roles. Before we proceed, we, like HP, introduce the notion of a causal setting which formally captures a scenario.

Definition 1. A (SC) *causal setting* is a triple $\langle \mathcal{D}, \sigma, \varphi(s) \rangle$ where $\mathcal{D}$ is a BAT, σ is a ground situation term such that $\mathcal{D} \models executable(\sigma)$, and $\varphi(s)$ is a SC formula uniform in s such that $\mathcal{D} \models \exists s(executable(s) \land \varphi(s))$.

Since the BAT $\mathcal{D}$ is fixed in our approach, we typically refer to $\langle \mathcal{D}, \sigma, \varphi(s) \rangle$ as just $\langle \sigma, \varphi(s) \rangle$.

4 The Achievement Causal Chain

Intuition provides few definite truths about actual causality, but we hold the following to be self-evident: If some action α of the action sequence σ triggers the formula $\varphi(s)$ to change its truth value from $false$ to $true$ relative to $\mathcal{D}$ and if there is no action in σ after α that changes the value of $\varphi(s)$ back to $false$, then α is an actual cause of achieving $\varphi(s)$ in σ. This statement is sound because (a) the narrative σ determines a total linear order on its actions, (b) change is associated with a particular element of that order, and (c) no change comes about other than by an action of σ. The next definition states this observation formally.

Definition 2. A causal setting $\mathcal{C} = \langle \sigma, \varphi(s) \rangle$ satisfies the *achievement condition* via the situation term $do(\alpha, \sigma') \sqsubseteq \sigma$ iff $\mathcal{D} \models \neg\varphi(\sigma') \land \forall s\, (do(\alpha, \sigma') \sqsubseteq s \sqsubseteq \sigma \rightarrow \varphi(s))$.

Whenever a causal setting $\mathcal{C}$ satisfies the achievement condition via $do(\alpha, \sigma')$, we say that the ground action α executed in σ' is a *(primary) achievement cause* in $\mathcal{C}$.

If a causal setting does not satisfy the achievement condition and $\varphi(s)$ is non-tautological and holds throughout the narrative σ, then we ascribe the achievement of $\varphi(s)$ to an unknowable cause masked by the initial situation S_0. If $\varphi(s)$ is a tautology, it legitimately has no cause. If $\varphi(\sigma)$ is not entailed by $\mathcal{D}$, then its achievement cause truly does not exist.

Example 2 (continued)**.** The entailment of Definition 2 holds when α is $tick$ and σ' is $do([c\_on,\ hi(d),\ tick,\ hi(e_1),\ hi(e_2)],\ S_0)$, meaning that the action #6 ($tick$ executed after σ') is the achievement cause of $Q(s)$ in σ.

The notion of the achievement condition forms our basic tool which, when used together with the single-step regression operator ρ, helps us not only find the single action that brings about the effect of interest, but also identify the actions that build up to it. Intuitively, $\rho[\varphi(s), \alpha]$ is the weakest precondition that must hold in a previous situation σ in order for $\varphi(s)$ to hold after performing α in σ. If we prove α to be an achievement cause of $\varphi(s)$ in $do(\alpha, \sigma)$, we can use regression ρ to obtain a formula that holds at σ and constitutes a necessary and sufficient condition for the achievement of $\varphi(s)$ via α. This new formula may have an achievement cause of its own which, by virtue of α, also constructively contributes to the achievement of $\varphi(s)$. By repeating this process, we can uncover the entire chain of actions that incrementally build up to the achievement of the ultimate effect. At the same time, we must not overlook the condition which makes the execution of α in σ even possible. This condition is conveniently captured by the right-hand side $\Pi_\alpha(s)$ of the precondition axiom for α and may have achievement causes of its own. The following inductive definition formalizes these intuitions.

Definition 3. If a causal setting $\mathcal{C} = \langle \sigma, \varphi(s) \rangle$ satisfies the achievement condition via some situation term $do(A(\bar{t}), \sigma') \sqsubseteq \sigma$ and α is an achievement cause in the causal setting $\langle \sigma',\ \rho[\varphi(s), A(\bar{t})] \land \Pi_A(\bar{t}, s) \rangle$, then α is an *achievement cause* in $\mathcal{C}$.

Clearly, the process of discovering intermediary achievement causes using single-step regression repeatedly cannot continue beyond S_0. Since the given narrative σ is a finite sequence, the achievement causes of $\mathcal{C}$ also form a finite sequence which we call the *achievement causal chain* of $\mathcal{C}$. Note that the actions of the achievement causal chain need not be adjacent in the action sequence of σ.

Example 3 (continued)**.** We found that the action $tick$ (#6) executed in $\sigma' = do([c\_on, hi(d), tick, hi(e_1), hi(e_2)], S_0)$ is the achievement cause of $Q(s)$. We can now use Definition 3 to find in σ the complete causal chain leading up to $Q(s)$. The one-step regression of $Q(s)$ through $tick$ is

$$\rho[Q(s), tick] = (\neg En(s) \lor High(d, s)) \land (En(s) \lor Q(s)).$$

Call $\psi(s)$ the conjunction of this formula and $ClockOn(s)$, the precondition of $tick$. By Definition 2, the achievement cause of $\psi(s)$ is the action $hi(e_1)$ executed in $do([c\_on,\ hi(d),\ tick],\ S_0)$. Therefore, $hi(e_1)$ is a secondary achievement cause of $Q(s)$. Applying Definition 3 again, we formulate another causal setting with the query

$$\rho[\psi(s), hi(e_1)] \land \Pi_{hi}(e_1, s) \equiv$$
$$High(d, s) \land ClockOn(s) \land \neg High(e_1, s)$$

and situation $do([c\_on, hi(d), tick], S_0)$, where $hi(d)$ is an achievement cause as a part of $do([c\_on, hi(d)], S_0)$. Regressing $High(d, s) \land ClockOn(s) \land \neg High(e_1, s)$ just past $hi(d)$, we obtain $\neg High(e_1, s) \land ClockOn(s)$, for which c_on is an achievement cause. Notice that the first action, c_on established preconditions for $tick$; were it not for c_on, $tick$ would have never happened! There are no more achievement causes of $Q(s)$ in σ aside from those already identified: c_on, $hi(d)$, $hi(e_1)$, $tick$. Observe that these are indeed the key events that lead to the achievement of $Q(s)$ in σ.

It is worth noting that our approach handled a classic instance of (late) preemption without appealing to contingencies occurring in neighbouring possible worlds, which is the essential strategy in counterfactual analyses. Namely, it correctly excluded $hi(e_2)$ from the causal chain for being preempted by $hi(e_1)$, although $hi(e_2)$ would have been sufficient, in the absence of $hi(e_1)$, for achieving $En(s)$ and $Q(s)$.

5 Maintenance Causes

The achievement causal chain explains precisely how a condition comes to be, but not how it persists throughout the remaining actions of the narrative. The narrative may well contain actions which could destroy the effect but were somehow neutralized. We formalize our intuitive understanding of protective actions using the notion of maintenance. Our general considerations are as follows. First, in a causal setting $\mathcal{C} = \langle \sigma, \varphi(s) \rangle$, if $\mathcal{D} \not\models \varphi(\sigma)$, then there is nothing to maintain. Therefore, $\mathcal{C}$ may have a maintenance cause only if $\mathcal{D} \models \varphi(\sigma)$. Second, every instance of maintenance involves at least two actions of σ, where one action—call it a *threat*—would falsify the goal φ were it not for the other action, the maintenance cause itself. Obviously, the maintenance cause must occur in σ before the corresponding threat. Third, if $\mathcal{C}$ satisfies the achievement condition via some $do(\alpha, \sigma')$, then neither α nor any action of σ' may be a threat to $\varphi(s)$, in accordance with the first consideration. If, alternatively, $\varphi(s)$ holds at S_0 and throughout σ, then any action of σ except the very first one may be a threat.

The key property of a threat is that it has the potential to falsify the effect (but did not do so in the narrative). A test for this property involves a construction of a hypothetical scenario where the suspected threat falsifies the effect. Such test is by nature counterfactual and, therefore, gives rise to the usual question: what alternative scenarios should we admit to the analysis? For the sake of generality, we require only that the alternative scenarios obey the rules of the world, and for the sake of tractability, that they do not contain too many actions. Both requirements are fulfilled by the following broad definition, where $len(s)$ is the number of actions in a situation term s and $len(S_0) = 0$.

Definition 4. A causal setting $\mathcal{C} = \langle \sigma, \varphi(s) \rangle$ satisfies the *maintenance condition* via a ground situation term $do(\tau, \sigma') \sqsubseteq \sigma$ iff $\sigma \neq S_0$ and $\mathcal{D} \models \forall s(\sigma' \sqsubseteq s \sqsubseteq \sigma \to \varphi(s))$ and $\mathcal{D} \models \exists s\big(\mathit{executable}(do(\tau, s)) \wedge \varphi(s) \wedge \neg\varphi(do(\tau, s)) \wedge len(do(\tau, s)) \leq len(\sigma)\big)$, in which case τ is a *threat* in $\mathcal{C}$.

A tighter definition of a threat would artificially decrease the search space of maintenance causes. If, through unchecked generality, we misdiagnose a harmless action as a threat, the subsequent achievement cause analysis would be unable to identify an action which neutralized the threat's harmful effects.

Before we define what is a maintenance cause, consider a threat τ to $\varphi(s)$. By the definition of regression, $\rho[\neg\varphi(s), \tau]$ is a formula that should hold to make sure $\varphi(s)$ becomes false after executing τ. Since we would like to preserve $\varphi(s)$, we are interested in the negation of this formula. But by the regression theorem, $\mathcal{D} \models \neg\rho[\neg\phi, \tau] \leftrightarrow \rho[\phi, \tau]$, so the formula expressing the maintenance goal is simply $\rho[\varphi(s), \tau]$. Notably, the set of achievement causes of this formula will include the achievement causes of $\varphi(s)$, because, intuitively, $\varphi(s)$ holds after τ in part due to being achieved.

Definition 5. Suppose a causal setting $\mathcal{C} = \langle \sigma, \varphi(s) \rangle$ satisfies the maintenance condition via some situation term $do(\tau, \sigma') \sqsubseteq \sigma$, where τ is a threat in $\mathcal{C}$. Let $\mathcal{C}'$ be the related causal setting $\langle \sigma', \rho[\varphi(s), \tau] \rangle$. If α is an achievement cause in $\mathcal{C}'$, we say that α is a *maintenance cause* in $\mathcal{C}$.

Example 4. Consider a formula $\psi(s)$ with quantifiers over object variables in the same BAT except that for the sake of example there is a countably infinite set of signal constants c_i for $i \geq 1$ with unique names. Let the query $\psi(s)$ be $\exists x \exists y(x \neq y \wedge High(x, s) \wedge High(y, s))$ — "there are at least two high signals". Let the actual situation σ be $do([hi(c_1), hi(c_2), hi(c_3), lo(c_1)], S_0)$.

By Definition 4, $lo(c_1)$ is a threat in this causal setting. By Definition 5, it yields a related causal setting with situation $do([hi(c_1),\ hi(c_2),\ hi(c_3)],\ S_0)$ and query $\rho[\psi(s), lo(c_1)]$, which simplifies to

$$\exists x \exists y(x \neq y \wedge High(x, s) \wedge High(y, s) \wedge x \neq c_1 \wedge y \neq c_1).$$

Applying Definition 2, we see this related causal setting has $hi(c_3)$ as an achievement cause. Therefore, the original causal setting has $hi(c_3)$ as a maintenance cause.

6 Actual Cause

The Definitions 2, 3, and 5 are centered around the top level of a given casual setting and fail to capture the interplay between achievement and maintenance causes at the deeper levels of analysis. Specifically, suppose a causal setting $\mathcal{C}'$ arises via the achievement (resp., maintenance) condition during the analysis of another setting $\mathcal{C}$. On its own, $\mathcal{C}'$ may have both achievement and maintenance causes, but, by Definition 3 (resp., 5), only the former are counted as causes of $\mathcal{C}$. On the natural assumption that all causes of a descendant setting are equally relevant to the ancestor setting, the following definition inductively combines all possible interactions between the achievement and maintenance conditions under the generic term *actual cause*.

Definition 6. Let α be a ground action and σ a narrative. We say that α is an *actual cause* in a causal setting $\mathcal{C} = \langle \sigma, \varphi(s) \rangle$ if at least one of the following conditions holds.

(a) $\mathcal{C}$ satisfies the achievement condition via $do(\alpha, \sigma') \sqsubseteq \sigma$.

(b) $\mathcal{C}$ satisfies the achievement condition via some situation term $do(A(\bar{t}), \sigma') \sqsubseteq \sigma$ and α is an actual cause in the causal setting $\langle \sigma', \rho[\varphi(s), A(\bar{t})] \wedge \Pi_A(\bar{t}, s) \rangle$.

(c) $\mathcal{C}$ satisfies the maintenance condition via $do(\tau, \sigma') \sqsubset \sigma$, and α is an actual cause in $\langle \sigma',\ \rho[\varphi(s), \tau] \rangle$.

Example 5 (continued from 4)**.** By Definition 6, the actions $hi(c_1)$, $hi(c_2)$, $hi(c_3)$ are all actual causes of $\psi(s)$. Notice that maintenance causes are just as important as achievement causes: the condition $\psi(s)$ was realized through the properties of objects c_1, c_2, but persevered by virtue of c_2, c_3. Achievement cause analysis alone disregards the role of c_3.

Example 6. Consider again our running example. By Definition 6, the 8-th action $lo(e_2)$ is a non-trivial actual cause of $Q(s)$ discovered through a combination of two maintenance condition. Intuitively, it is causally important because it disables the flop-flop, preventing the actions $tick$ (#11) and $lo(d)$ (#10) from destroying $Q(s)$ — both are threats in their respective settings.

7 The Halpern-Pearl Approach

Halpern and Pearl (2005), following the motivation of [Lewis, 1974], base their formal account of actual causality on the notion of a counterfactual — a conditional statement whose premise is contrary to fact. They construct counterfactual statements in a formal language whose semantics is defined relative to a *causal setting* (see below). A *causal model* M is a tuple $\langle \mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F} \rangle$, where $\mathcal{U}$ and $\mathcal{V}$ are finite disjoint sets of *exogenous* and *endogenous* variables, respectively, with each variable taking various values from an underlying domain. The function $\mathcal{R}$ maps every variable $Z \in \mathcal{U} \cup \mathcal{V}$ to a non-empty set $\mathcal{R}(Z)$ of possible values. $\mathcal{F}$ is a set of total functions $\{F_X : \times_{Z \in \mathcal{U} \cup \mathcal{V} \setminus \{X\}} \mathcal{R}(Z) \mapsto \mathcal{R}(X) \mid X \in \mathcal{V}\}$ which act like structural equations; each tuple of values assigned to the variables (excluding X) maps to a single value of X. Intuitively, for each endogenous variable X, F_X encodes the entirety of causal laws which determine X by mapping every value assignment on all variables except X to some value of X. The values of exogenous variables $\mathcal{U}$ are set externally; a tuple $\bar{V}_U$ of values for $\mathcal{U}$ is called a *context* of M, and the pair $(M, \bar{V}_U)$ constitutes a *causal setting*. The tuple $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ is called the *signature* of M. The set of functions $\mathcal{F}$ determines a partial *dependency order* $X \preceq Y$ on endogenous variables X, Y. Namely, Y depends on X, $X \preceq Y$, if either X affects Y directly by virtue of F_Y, or indirectly via intermediate functions. It is ubiquitously assumed that a given causal model is *acyclic*, that is, for each context $\bar{V}_U$ of M, there is a partial order $\preceq$ on $\mathcal{V}$ that is anti-symmetric, reflexive and transitive. This assumption guarantees the existence of a unique solution to the equations $\mathcal{F}$.

The language of the HP approach is as follows. A primitive *event* is a formula $X = V_X$ where $X \in \mathcal{V}$ and $V_X \in \mathcal{R}(X)$. We call a Boolean combination of primitive events a *HP query*. A general *causal formula* is one of the form $[Y_1 \leftarrow V_{Y_1}, \ldots, Y_k \leftarrow V_{Y_k}]\phi$ where ϕ is a HP query, Y_i for $1 \leq i \leq k$ are distinct variables from $\mathcal{V}$, and $V_{Y_i} \in \mathcal{R}(Y_i)$. (We abbreviate $[Y_1 \leftarrow V_{Y_1}, \ldots, Y_k \leftarrow V_{Y_k}]$ as $[\bar{Y} \leftarrow \bar{V}_Y]$ and call it an *intervention*.) A primitive event $X = V_X$ is satisfied in a causal setting $(M, \bar{V}_U)$, denoted $(M, \bar{V}_U) \models (X = V_X)$, if X takes on the value V_X in the unique solution to the equations $\mathcal{F}$ once $\mathcal{U}$ are set to $\bar{V}_U$. HP queries are interpreted following the usual rules for Boolean connectives. Finally, $(M, \bar{V}_U) \models [Y_1 \leftarrow V_{Y_1}, \ldots, Y_k \leftarrow V_{Y_k}]\phi$ iff $(M', \bar{V}_U) \models \phi$ where M' is obtained from M by replacing each $F_{Y_i} \in \mathcal{F}$ by the trivial function $F_{Y_i} : \times_{Z \in \mathcal{U} \cup \mathcal{V} \setminus \{X\}} \mathcal{R}(Z) \mapsto V_{Y_i}$ that fixes Y_i to a constant V_{Y_i} for all the values of arguments.

In this paper, we focus on the so-called *modified* HP definition, or HP^m, of actual cause [Hopkins, 2005; Halpern, 2015; 2016] because it is the most recent, intuitively appealing, and thoroughly connected with older definitions by formal results in [Halpern, 2016]. According to this definition, the conjunction of primitive events $\bar{X} = \bar{V}_X$ (short for $X_1 = V_{X_1} \land \ldots \land X_k = V_{X_k}$) is an *actual cause* in $(M, \bar{V}_U)$ of a HP query ϕ if all of the following conditions hold:

1. $(M, \bar{V}_U) \models (\bar{X} = \bar{V}_X)$ and $(M, \bar{V}_U) \models \phi$.
2. There exists a set $\bar{W}$ (disjoint from $\bar{X}$) of variables in $\mathcal{V}$ with $(M, \bar{V}_U) \models (\bar{W} = \bar{V}_W)$ and a setting $\bar{V}'_X$ of variables $\bar{X}$ such that $(M, \bar{V}_U) \models [\bar{X} \leftarrow \bar{V}'_X, \bar{W} \leftarrow \bar{V}_W]\neg\phi$.
3. No proper sub-conjunction of $(\bar{X} = \bar{V}_X)$ satisfies 1, 2.

Example 7. Consider the two well-known "Forest Fire" examples from [Halpern and Pearl, 2005; Halpern, 2016]. Both have the same set of endogenous variables: MD (match dropped by arsonist), L (lightning strike), *FF* (forest is on fire). In both cases, MD and L are set to $true$ by the context. The model M_d for the *disjunctive* scenario has it that either one of the events $(MD = true)$, $(L = true)$ is sufficient to start a fire, so the equation for *FF* is *FF* $:= (MD = true) \lor (L = true)$. The model M_c for the *conjunctive* scenario requires both events in order to create a forest fire, so *FF* $:= (MD = true) \land (L = true)$. By HP^m, neither $(MD = true)$ nor $(L = true)$ are singleton actual causes in M_d because it is impossible to fulfill part 2 of the definition above by setting either variable to $false$, but the conjunction $(MD = true) \land (L = true)$ is deemed an actual cause. In contrast, in M_c, both $(MD = true)$ and $(L = true)$ are singleton actual causes because setting one of $\{MD, L\}$ to $false$ makes the forest fire impossible, but their conjunction is not an actual cause because it violates the minimality condition.

8 Formal Relationship with HP

We establish a common ground between the two formalisms by axiomatizing causal models in SC.

Let $(M, \bar{V}_U)$ be a HP causal setting where $M = \langle \mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F} \rangle$ is an acyclic causal model and $\bar{V}_U$ a context. We assume that $\mathcal{U}$, $\mathcal{V}$, and the range of $\mathcal{R}$ are finite sets and there are no collisions between constants for variable and value symbols.

We construct a BAT $\mathcal{D}$ from $(M, \bar{V}_U)$ as follows. We treat $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{R}(X)$ for all $X \in \mathcal{U} \cup \mathcal{V}$ as sets of SC constant symbols for which we introduce unique name axioms. If $S = \{C_1, \ldots, C_n\}$ is a set of constants and y is a SC object term, the expression $y \in S$ denotes $(y = C_1 \lor \ldots \lor y = C_n)$. If $X \in \mathcal{U} \cup \mathcal{V}$ with $\mathcal{R}(X) = \{V_1, \ldots, V_n\}$, $y \in \mathcal{R}(X)$ denotes $(y = V_1 \lor \ldots \lor y = V_n)$. To represent functions $\mathcal{F}$, we introduce a situation-independent relational symbol f with arity $1 + |\mathcal{U} \cup \mathcal{V}| + 1$ where the first argument is the name of the variable (X) which $F_X \in \mathcal{F}$ determines, the last argument is the value which F_X assigns to X, and the arguments in between are the values of variables $\mathcal{U} \cup \mathcal{V}$ arranged in some predetermined order. The actions of $\mathcal{D}$ are $get(x, v)$, meaning *compute the value of the endogenous variable x using $F_x \in \mathcal{F}$*, and $set(x, v)$, meaning *ignore F_x and force the value v upon x*. The only fluent of $\mathcal{D}$ is the relational fluent $V(x, v, s)$ stating that v is the value of the endogenous variable x in situation s.

Let $Det(x, v, s)$ be an abbreviation for

$$\forall v_1 \ldots \forall v_N \,.\, \bigwedge_{1 \leq i \leq N} \exists y \big\{ y = Z_i \land v_i \in \mathcal{R}(Z_i) \land \forall v' (V(y, v', s) \rightarrow v_i = v') \big\} \rightarrow f(x, v_1, \ldots, v_N, v),$$

where $\mathcal{U} \cup \mathcal{V} = \{Z_1, \ldots, Z_N\}$. $Det(x, v, s)$ means that the value of variable x is *determined* in s to be v. $Det(x, v, s)$ holds true when the values v_i which exist in s, when bound to appropriate arguments of f, unequivocally assign v to x. This means, crucially, that x may be determined as soon as some—but not necessarily all—of the variables on which it "depends" (as per $\preceq$) have acquired values.

The axioms of $\mathcal{D}$ are as follows.

$$\begin{aligned}
&\textstyle\bigwedge_{X\in\mathcal{V}} \neg\exists v(V(X,v,S_0)), \quad \bigwedge_{V_Y\in\bar{V}_U} V(Y,V_Y,S_0),\\
&Poss(set(x,v),s) \leftrightarrow\\
&\quad \textstyle\bigvee_{X\in\mathcal{V}}(x = X \wedge v \in \mathcal{R}(X)) \wedge \neg\exists v'\, V(x,v',s),\\
&Poss(get(x,v),s) \leftrightarrow\\
&\quad x \in \mathcal{V} \wedge \neg\exists v'\, V(x,v',s) \wedge Det(x,v,s),\\
&V(x,v,do(a,s)) \leftrightarrow\\
&\quad a = get(x,v) \vee a = set(x,v) \vee V(x,v,s).
\end{aligned}$$

In words, none of the endogenous variables have values at S_0, and all exogenous variables have values at S_0 as specified by the context. It is possible to force a value v upon x as long as x is an endogenous variable, v is in the range of x, and x has not yet acquired a value. It is possible to compute the value of x as long as x is an endogenous variable which has not yet acquired a value but which is destined at s to get the value v. Overall, the theory models all possible propagations of values (including interventions) throughout the set of variables according to the structural equations. As we are interested only in those situations where all variables have acquired values, which represent a unique solution to $\mathcal{F}$, we introduce an abbreviation $terminal(s)$ for the expression $executable(s) \wedge \neg\exists a(Poss(a,s))$. In order to refer to situations under specific interventions, we use the abbreviation $interv_{Y_1\leftarrow V_{Y_1},\ldots,Y_k\leftarrow V_{Y_k}}(s)$ which stands for $terminal(s) \wedge \forall x\forall v.[\exists s'(do(set(x,v),s') \sqsubseteq s) \leftrightarrow \bigvee_{1\leq i\leq k}(x = Y_i \wedge v = V_{Y_i})]$. The special case $interv_{\emptyset}(s)$ describes s under an empty intervention.

Finally, given a HP query ϕ, we obtain a corresponding SC query $\hat{\phi}$ from ϕ by replacing each primitive event $(X = V_X)$ by $V(X,V_X,s)$. Thus, $\hat{\phi}$ is ground in all object arguments and uniform in s. It is tedious but straightforward to prove the correctness of our translation relative to a HP causal setting.

Theorem 1. *Let $(M,\bar{V}_U)$ be a HP causal setting, $[\bar{Y} \leftarrow \bar{V}_Y]\phi$ an arbitrary causal formula over M, and $\mathcal{D}$ a BAT obtained from $(M,\bar{V}_U)$. Then $(M,\bar{V}_U) \models [\bar{Y} \leftarrow \bar{V}_Y]\phi$ iff $\mathcal{D} \models (\forall s).\ interv_{\bar{Y}\leftarrow\bar{V}_Y}(s) \rightarrow \hat{\phi}(s)$.*

With this result, we can easily translate HP^m to the language of SC and formally compare the two approaches.

Theorem 2. *Let $(M,\bar{V}_U)$ be a HP causal setting and ϕ a HP query over M. Let $\mathcal{D}$ be a BAT obtained from $(M,\bar{V}_U)$ as described above. Let $X \in \mathcal{V}$ and $V_X \in \mathcal{R}(X)$.*

1. *$(X = V_X)$ is a singleton cause of ϕ in $(M,\bar{V}_U)$ according to HP^m if and only if $get(X,V_X) \in \sigma$ appears in the achievement causal chain of $\langle\sigma,\hat{\phi}(s)\rangle$ for every ground situation term σ of $\mathcal{D}$ such that $\mathcal{D} \models interv_{\emptyset}(\sigma)$.*
2. *$(X = V_X)$ is a part of a cause of ϕ in $(M,\bar{V}_U)$ according to HP^m if and only if there exists a ground situation term σ of $\mathcal{D}$ such that $\mathcal{D} \models interv_{\emptyset}(\sigma)$ and $get(X,V_X) \in \sigma$ appears in the achievement causal chain of $\langle\sigma,\hat{\phi}(s)\rangle$.*

The proof of Theorem 2 is quite involved and is not shown due to lack of space. By an immediate corollary, achievement cause analysis alone captures all HP^m causes.

Example 8. (cont.) Consider a translation of the disjunctive Forest Fire causal model M_d. The corresponding terminal narratives σ are

$$\begin{aligned}
&do([get(MD,true), get(L,true), get(FF,true)], S_0),\\
&do([get(L,true), get(MD,true), get(FF,true)], S_0),\\
&do([get(MD,true), get(FF,true), get(L,true)], S_0),\\
&do([get(L,true), get(FF,true), get(MD,true)], S_0).
\end{aligned}$$

Action $get(MD,true)$ is a part of the causal chain of $\langle\sigma, V(FF,true,s)\rangle$ only for the first and third choice of σ. Similarly, $get(L,true)$ is an achievement cause only for the second and fourth choice. By Part 1 of Theorem 2, they are not actual causes according to HP^m. By Part 2 of Theorem 2, they are both parts of an actual cause according to HP^m. This agrees with conclusions of the original HP causal model.

9 Discussion

Our approach shifts the focus away from causal models and towards first-order logic representation of the underlying dynamics of the scenario. There are other attempts to step away from purely counterfactual analysis [Vennekens *et al.*, 2010; Vennekens, 2011; Beckers and Vennekens, 2012; 2016], but they share the same expressivity limitations. Curiously, [Vennekens *et al.*, 2010] consider SC to be too expressive, stating that "SC contains many features that go beyond what is traditionally expressed in a causal model. For typical causal reasoning problems, these features are not needed". To refute this and to see where we stand with respect to other approaches, let us consider three telling examples featured in [Beckers and Vennekens, 2012; 2016] and discussed in other papers. Assume all fluents are false at S_0.

Example 9. *Assassin poisons victim's coffee, victim drinks it and dies. If assassin had not poisoned the coffee, his backup would have, and victim would still have died.*

This example from [Hitchcock, 2007] illustrates *early preemption*, namely that the causal link from the backup to victim's death is preempted by the assassin. Let the actions be $assassin$ and $backup$ (the two acts of poisoning the coffee) and $drink$. Let the fluents be $P(s)$ meaning "coffee contains poison" and $D(s)$ meaning "the victim is dead".

$$\begin{aligned}
&Poss(assassin,s), Poss(backup,s)\\
&Poss(drink,s) \leftrightarrow P(s),\\
&P(do(a,s)) \leftrightarrow a = assassin \vee a = backup \vee P(s),\\
&D(do(a,s)) \leftrightarrow [a = drink \wedge P(s)] \vee D(s).
\end{aligned}$$

The narrative is $\sigma = do([assassin, drink], S_0)$. By our analysis, all of σ is an achievement causal chain. This agrees with HP and [Hitchcock, 2007] but disagrees with Beckers and Vennekens who believe that $assassin$ is not an actual cause. Rather than appeal to intuition, we just point out that the causal roles assumed by the assassin and his backup are clearly distinct *in the given scenario*.

Example 10. *An engineer is standing by a switch in the railroad track. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track instead of the right. Since the tracks re-converge up ahead, the train arrives at its destination all the same.*

This example from by [Paul and Hall, 2013] illustrates the distinction between causation and determination. Beckers and Vennekens state that it is isomorphic to the previous one, while the intuition about its causes is the polar opposite. In fact, the two examples are isomorphic only within the expressivity bounds of causal models and CP-logic.

Let the fluent $In(s)$ mean that the train is on the section of the track leading to the first junction, let $L(s)$ (resp., $R(s)$) mean that it is on the left-hand track (resp., right), and let $Out(s)$ mean that it is on the section of the track past the second junction. Let the fluent $Sw(s)$ mean that the switch is engaged and $Arrived(s)$ that the train has arrived. Let the actions be $flip$ (engineer flips the switch), $fork_1$ (train passes first junction), $fork_2$ (train passes second junction), and $arrive$ (self-explanatory). Let only $In(s)$ hold at S_0.

$$\begin{gathered}
Poss(flip, s),\ \ Poss(fork_1, s) \leftrightarrow In(s), \\
Poss(fork_2,s) \leftrightarrow L(s) \vee R(s), Poss(arrive,s) \leftrightarrow Out(s), \\
In(do(a, s)) \leftrightarrow In(s) \wedge a \neq fork_1, \\
L(do(a, s)) \leftrightarrow a{=}fork_1 \wedge Sw(s) \vee L(s) \wedge a \neq fork_2, \\
R(do(a, s)) \leftrightarrow a{=}fork_1 \wedge \neg Sw(s) \vee R(s) \wedge a \neq fork_2, \\
Out(do(a, s)) \leftrightarrow a{=}fork_2 \vee Out(s), \\
Sw(do(a, s)) \leftrightarrow a{=}flip \vee Sw(s) \wedge a \neq flip, \\
Arrived(do(a, s)) \leftrightarrow a{=}arrive \vee Arrived(s).
\end{gathered}$$

The narrative σ is $do([flip,fork_1,fork_2, arrive], S_0)$. By our analysis, the $flip$ action is not an actual cause of train's arrival. This conclusion is elaboration tolerant [McCarthy, 1987] as long as the relation between L, R, Sw is preserved. For HP, the answer depends on how model is constructed and which definition is applied. [Pearl, 2000] calls this class of problems "switching causation" and argues that flipping switch is a cause (see Section 10.3.4, p.324-5). Both [Pearl, 2000] and [Halpern and Pearl, 2005] argue that switch is a cause, while, according to HP^m, it is not.

Example 11. *Assistant Bodyguard puts a harmless antidote in victim's coffee. Buddy who knows about the antidote poisons the coffee; he would not have done so otherwise. Victim drinks the coffee and survives.*

This example is called "Careful Poisoning" in [Weslake, 2013] and left as a challenge for future work. Let the actions be $antidote, poison, drink$. The fluents $P(s)$, $D(s)$ are as before, and the fluent $A(s)$ means "coffee contains antidote".

$$\begin{gathered}
Poss(antidote, s), Poss(drink, s), \\
Poss(poison, s) \leftrightarrow A(s), \\
A(do(a, s)) \leftrightarrow a = antidote \vee A(s), \\
P(do(a, s)) \leftrightarrow a = poison \vee P(s), \\
D(do(a, s)) \leftrightarrow [a = drink \wedge P(s) \wedge \neg A(s)] \vee D(s).
\end{gathered}$$

$\sigma = do([antidote, poison, drink], S_0)$, so $\mathcal{D} \models \neg D(\sigma)$. In fact, $\neg D(s)$ holds throughout the narrative, so it has no achievement causes. It has no maintenance causes either: $drink$ is a threat to $\neg D(s)$, yielding a new causal setting $\langle do([antidote, poison], S_0), \neg D(s) \wedge (\neg P(s) \vee A(s)\,)\,\rangle$ with no achievement causes. The action $poison$ could be a threat, but it does not qualify as such by our definition: no executable situation admits $poison$ in the absence of the antidote, owing to the precondition for $poison$. Therefore, the given causal setting contains no causes. This agrees with Beckers and Vennekens and disagrees with Hitchcock and HP.

There exist multiple examples where the results of the HP approach cannot be reconciled with intuitive understanding—which, incidentally, the approach treats as the only measure of merit. This problem was traced by [Hopkins and Pearl, 2007] and [Glymour *et al.*, 2010] to the limited expressiveness of causal models. Causal models do not distinguish between enduring conditions and transitions and cannot to model the absence of an event except as the presence of its opposite; examples where this leads to absurd conclusions are easy to come by, see *e.g.* [Hopkins and Pearl, 2007]. An explicit notion of an action solves these problems.

Addressing the lack of expressivity, [Hopkins and Pearl, 2007] re-defined causal models in the language of SC, but they preserved the implicit possible worlds semantics of causal formulas and dropped the requirement that situations be executable. The latter is especially problematic, since dismissing preconditions results in paradoxes and makes inferences untrustworthy. Our work reaps the benefits which [Hopkins and Pearl, 2007] aimed at but does not suffer from the issues associated with giving a meaningful definition of a counterfactual in SC, which appears to be no easy task. A counterfactual query not relativized to a particular scenario can be formulated in SC without special tools [Lin and Soutchanski, 2011], but it is not clear how such queries can be useful for defining actual causality. An original study conducted in [Costello and McCarthy, 1999] perhaps comes closest to a good definition of a counterfactual in SC, but it operates outside of the well-studied basic action theories and it is not concerned with actual causality. There exist numerous studies of the semantics of causal models and the relationship of causal models to various logics, such as an elaborate axiomatization of causal models [Halpern, 2000] and a logical representation [Bochman and Lifschitz, 2015] of causal models in a non-monotonic logic which encompasses general causation as a foundational principle. The approach of [Finzi and Lukasiewicz, 2003] combines causal models with independent choice logic. Finally, there are methodological or technical critiques of the causal model approach, exemplified by [Glymour *et al.*, 2010], [Menzies, 2014], [Livengood, 2013], [Weslake, 2013] and [Baumgartner, 2013].

It is clear that a more broad definition of actual cause requires more expressive action theories that can model not only sequences of actions, but can also include explicit time and concurrent actions. Only after that one can try to analyze some of the popular examples of actual causation formulated in philosophical literature; some of those examples sound deceptively simple, but faithful modelling of them requires time, concurrency and natural actions [Reiter, 2001]. This does not imply that future research should focus only on popular scenarios proposed by philosophers. To the contrary, we firmly believe that the future of causal research is in elaborating computational methodology for the analysis of complex technical systems.

Acknowledgement: We thank the Natural Sciences and Engineering Research Council of Canada for financial support.

References

[Baumgartner, 2013] Michael Baumgartner. A regularity theoretic approach to actual causation. *Erkenntnis*, 78(Supplement 1 "Actual Causation"):85–109, 2013.

[Beckers and Vennekens, 2012] Sander Beckers and Joost Vennekens. Counterfactual dependency and actual causation in CP-logic and structural models: a comparison. In *Proceedings of the Sixth Starting AI Researchers Symposium*, volume 241, pages 35–46, 2012.

[Beckers and Vennekens, 2016] Sander Beckers and Joost Vennekens. A principled approach to defining actual causation. *Synthese*, Oct 2016.

[Bochman and Lifschitz, 2015] Alexander Bochman and Vladimir Lifschitz. Pearl's causality in a logical setting. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 1446–1452. AAAI Press, 2015.

[Costello and McCarthy, 1999] Tom Costello and John McCarthy. Useful counterfactuals. *Electron. Trans. Artif. Intell.*, 3(A):51–76, 1999.

[Eiter and Lukasiewicz, 2002] Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artif. Intell.*, 142(1):53–89, 2002.

[Finzi and Lukasiewicz, 2003] Alberto Finzi and Thomas Lukasiewicz. Structure-based causes and explanations in the independent choice logic. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 225–323, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[Glymour *et al.*, 2010] Clark Glymour, David Danks, Bruce Glymour, Frederick Eberhardt, Joseph Ramsey, Richard Scheines, Peter Spirtes, Choh Man Teng, and Jiji Zhang. Actual causation: a stone soup essay. *Synthese*, 175(2):169–192, 2010.

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

[Halpern, 2000] Joseph Y. Halpern. Axiomatizing causal reasoning. *J. Artif. Intell. Res. (JAIR)*, 12:317–337, 2000.

[Halpern, 2015] Joseph Y Halpern. A modification of the Halpern-Pearl definition of causality. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3022–3033, 2015.

[Halpern, 2016] Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2016. ISBN 9780262035026.

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

[Hopkins and Pearl, 2007] Mark Hopkins and Judea Pearl. Causality and counterfactuals in the situation calculus. *Journal of Logic and Computation*, 17(5):939–953, 2007.

[Hopkins, 2005] Mark Hopkins. *The Actual Cause: From Intuition to Automation*. PhD thesis, University of California Los Angeles, 2005.

[Lewis, 1974] David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1974.

[Lin and Soutchanski, 2011] Fangzhen Lin and Mikhail Soutchanski. Causal theories of actions revisited. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[Livengood, 2013] Jonathan Livengood. Actual causation and simple voting scenarios. *Nous*, 47(2):316–345, 2013.

[McCarthy and Hayes, 1969] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Readings in artificial intelligence*, pages 431–450, 1969.

[McCarthy, 1987] John McCarthy. Generality in artificial intelligence. *Commun. ACM*, 30(12):1029–1035, 1987.

[Menzies, 2014] Peter Menzies. Counterfactual theories of causation. In *Stanford Encyclopedia of Philosophy*, `https://plato.stanford.edu/entries/causation-counterfactual/`, 2014. Retrieved on January 15, 2017.

[Paul and Hall, 2013] L.A. Paul and Ned Hall. *Causation: a user's guide*. Oxford University Press, ISBN 978-0199673452, 2013.

[Pearl, 1998] Judea Pearl. On the definition of actual cause. Technical report, R-259, University of California Los Angeles, 1998.

[Pearl, 2000] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1 edition, 2000.

[Reiter, 1991] Raymond Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. *Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy*, 27:359–380, 1991.

[Reiter, 2001] Raymond Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT press Cambridge, 2001.

[Simon, 1977] Herbert A Simon. Causal ordering and identifiability. In *Models of Discovery*, pages 53–80. Springer, 1977.

[Vennekens *et al.*, 2010] Joost Vennekens, Maurice Bruynooghe, and Marc Denecker. Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In *European Workshop on Logics in Artificial Intelligence*, pages 313–325. Springer, 2010.

[Vennekens, 2011] Joost Vennekens. Actual causation in CP-logic. *TPLP*, 11(4-5):647–662, 2011. `http://arxiv.org/abs/1107.4865`.

[Weslake, 2013] Brad Weslake. *A Partial Theory of Actual Causation*. `http://bweslake.s3.amazonaws.com/research/papers/weslake_ac.pdf`, 2013. Version c4eb488. Retrieved on July 18, 2017.