The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Situation Calculus Semantics for Actual Causality

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

Mikhail Soutchanski
Ryerson University
Toronto, Canada
http://www.scs.ryerson.ca/mes

Abstract

The definitions of actual cause given by Pearl and Halpern (HP) in the framework of causal models provided vital computational insight into an old philosophical problem but by no means resolved it. One source of concern is the lack of objective criteria for selecting possible worlds to be admitted into the counterfactual analysis, epitomized by the competition between multiple proposals by HP and others. Another concern is due to the modest expressivity of propositional-level structural equations which limits their applicability and, arguably, contributes to the the former problem. We tackle both of these issues using a novel approach. We build our definition of actual cause from first principles in the context of atemporal situation calculus (SC) action theories with sequential actions. As a result, we can successfully identify actual causes of conditions expressed in first-order logic. We validate the HP approach by providing a formal translation from causal models to SC and proving a relationship between our definitions of actual cause and that of HP. Using well-known and new examples, we show that long-standing disagreements between alternative definitions of actual causality can be mitigated by faithful SC modelling of the domains.

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 enquiry 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 1953; 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 (Halpern 2015). 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. Thus, causal models are too coarse to distinguish between enduring conditions and transitional events, providing only atomic propositions to model both. Moreover, causal models represent presence and absence of an event identically — by assigning a value to a propositional variable. Both of these deficiencies stem from the lack of a mechanism for modelling change over time.

Since counterfactual 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 the situation calculus 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.

We start with a brief introduction to situation calculus (SC) in Section 2, and to HP's causal models in Section 3. In Section 4 we propose our new definition of an achievement cause. We investigate the formal relation between our definition and the recent HP's ("modified") definition of an actual cause in Section 5. Finally, in Section 6 we discuss related work, and then conclude in Section 7.

2 Situation Calculus

SC is proposed in (McCarthy and Hayes 1969) and elaborated in (Reiter 2001). In the Reiter's SC, 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 $s_1 \sqsubset s_2 \vee s_1 = s_2$. 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 $Poss(A(\vec{x}), s) \leftrightarrow \Pi_A(\vec{x}, s)$, where all free variables are implicitly $\forall$-quantified, and $\Pi(\vec{x}, s)$ 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)

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

where the fluent predicate $F(\vec{x}, s)$ represents a situation-dependent relation over a tuple of objects $\vec{x}$, uniform formulas $\psi^+(\vec{x}, a, s)$ and $\psi^-(\vec{x}, a, s)$ 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 abbreviation $executable(s)$ means that each action mentioned in the situation term s was possible in the situation in which it was executed:

$$executable(s) \stackrel{\text{def}}{=} \forall a \forall s'(do(a, s') \sqsubseteq s \rightarrow Poss(a, s')).$$

In SC, the formulas $\forall s\, \psi(s)$, where $\psi(s)$ is uniform in s, are called state constraints since they represent conditions true in every state (Lin and Reiter 1994). Without loss of generality, we would assume that any given BAT $\mathcal{D}$ entails all state constraints, *i.e.*, they are compiled into $\mathcal{D}$. The basic computational challenge, called the *projection problem*, is the task of establishing whether a BAT entails, for an executable ground situation term σ, that a query sentence $\phi(\sigma)$ holds, where $\phi(\sigma)$ is a formula uniform in σ. This problem can be solved using the one-step regression operator ρ. The expression $\rho[\varphi(s), \alpha]$ denotes the formula obtained from $\varphi(s)$ by replacing each fluent atom $F(\vec{t}, s)$ that occurs in φ with the right-hand side of the SSA for F, where an action variable a is replaced with the ground action term α, and then the resulting formula is simplified using unique name axioms for actions and constants. Similarly to the theorem about multi-step regression $\mathcal{R}$ 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 $\mathcal{D} \models \varphi(do(\alpha, s)) \leftrightarrow \rho[\varphi(s), \alpha]$.

3 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 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 finite 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 subsequently 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 *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. Since M is acyclic, M' remains acyclic too.

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 $X_1 = V_{X_1} \wedge \ldots \wedge X_k = V_{X_k}$) is an *actual cause* in $(M, \bar{V}_U)$ of a HP query ϕ if all 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.

Notice that in Item 2, according to $(M, \bar{V}_U) \models (\bar{W} = \bar{V}_W)$, interventions that set variables in $\bar{X}$ to counterfactual values $\bar{V}'_X$ have to set all variables in $\bar{W}$ to their actual values $\bar{V}_W$ in the actual context. The tuple $\langle \bar{W}, \bar{V}_W, \bar{V}'_X \rangle$ is called a *witness* to the fact that $(\bar{X} = \bar{V}_X)$ is a cause of ϕ.

Example 1. 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 becomes $FF := (MD=true) \vee (L=true)$. The model M_c for the *conjunctive* scenario requires both events in order to create a forest fire, so $FF := (MD{=}true) \wedge (L{=}true)$. 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 $(MD = true) \wedge (L = true)$ 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 minimality (condition 3).

HP^m is an improvement over the original proposal by (Halpern and Pearl 2005), in part, because it is able to differentiate between such conjunctive and disjunctive scenarios.

4 Proposal: The Achievement Causal Chain

We propose to axiomatize a dynamic world using a situation calculus theory and derive actual causality from first principles. Specifically, to represent a "scenario", we consider a BAT $\mathcal{D}$ and an accompanying narrative describing the actions/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. To formally capture a scenario, we, like HP, introduce the notion of a causal setting.

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 situation calculus formula uniform in s such that $\mathcal{D} \models \exists s(executable(s) \wedge \varphi(s))$.

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

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 $\mathcal{D} \models \neg\varphi(\sigma') \wedge \forall s\,(do(\alpha, \sigma') \sqsubseteq s \sqsubseteq \sigma \rightarrow \varphi(s))$.

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 $\mathcal{D}$, meaning that $\varphi(s)$ is not achieved by the end of the narrative, then its achievement cause truly does not exist.

Example 1 (cont.)**.** We axiomatize the conjunctive Forest Fire example in a straight-forward way. Let $MD(s)$, $L(s)$, $FF(s)$ be fluents; let md, l be the agent's actions which affect the respective fluents, and let ff be a (natural) event triggered by the previous actions.

$$Poss(md, s), Poss(l, s), Poss(ff, s) \leftrightarrow MD(s) \wedge L(s),$$
$$MD(do(a, s)) \leftrightarrow a{=}md \vee MD(s),$$
$$L(do(a, s)) \leftrightarrow a{=}l \vee L(s),$$
$$FF(do(a, s)) \leftrightarrow a{=}ff \vee FF(s).$$

The story does not specify a temporal order between md and l, so w.l.o.g. let us fix a narrative where the match is dropped before the lightning strike: $\sigma = do([md, l, ff], S_0)$. The causal setting $\langle \sigma, FF(s) \rangle$ satisfies the achievement condition via the event ff, so ff executed in $do([md, l], S_0)$ is an achievement cause. This is obviously true, but not very useful. To find the root cause we need a deeper analysis.

The notion of the achievement condition mentioned before 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 single-step 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 of the precondition axiom for α and may have achievement causes of its own. To sum up: if α is an achievement cause of $\varphi(s)$ in $do(\alpha, \sigma')$, then $\rho[\varphi(s), \alpha]$ and the precondition $\Pi_\alpha(s)$ of α, taken together, express the condition which (a) holds at σ', (b) is necessary and sufficient for executing α in σ', and (c) is necessary and sufficient for achieving $\varphi(s)$ via α. The following inductive definition formalizes this intuition.

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 $\langle \sigma', \rho[\varphi(s), A(\bar{t})] \wedge \Pi_A(\bar{t}, s) \rangle$, 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. In fact, in σ they can be interspersed with other actions irrelevant to the achievement of φ.

Example 1 (cont.)**.** Computing $\rho[FF(s), ff] \wedge Poss(ff, s)$ by Definition 3 gives rise to a new causal setting $\langle do([md, l], S_0), MD(s) \wedge L(s) \rangle$. This setting satisfies the achievement condition via the action l, so l executed in $do([md], S_0)$ is an achievement cause. This yields yet another setting $\langle do([md], S_0), MD(s) \rangle$ which meets the achievement condition via md, and the analysis terminates. Thus, in this example, all actions of σ constitute a causal chain leading up to $FF(s)$. Observe that our choice of the narrative $do([md, l, ff], S_0)$ over the other possibility $do([l, md, ff], S_0)$ does not affect the conclusion: in the alternative narrative, all actions are also deemed causes.

To model disjunctive Forest Fire, we replace the precondition axiom for ff by $Poss(ff, s) \leftrightarrow MD(s) \vee L(s)$. Like before, the causal setting $\langle do([md, l, ff], S_0), FF(s) \rangle$ has an achievement cause ff and generates another setting $\langle do([md, l], S_0), MD(s) \vee L(s) \rangle$. In contrast to the conjunctive case, however, this new setting has md as an achievement cause, and the analysis terminates at S_0. The complete causal chain here consists of md, ff. The lightning strike is overlooked because the match was sufficient for starting a fire and occurred first. This may seem like a limitation of our approach, but consider the alternative narrative $do([l, md, ff], S_0)$: there, the causal chain is l, ff, so we are able to identify all causally relevant events by considering all possible narratives that fit the story. This example also illustrates just how well our approach handles preemption: if the story had stipulated that the match was dropped before the lightning strike, we would automatically discount the lightning strike as a cause without having to construct elaborate contingencies along the lines of HP.

Note that our axiomatization of Forest Fire contains a triggered event ff. To provide SC semantics to such events, we adhere to Reiter's notion of *natural actions* (Reiter 2001) which modifies $executable(s)$ to force natural actions to occur as soon as their preconditions are realized.

5 Formal Relationship with HP

In order to prove a formal relationship between Definition 3 and HP^m, we first need to establish a common ground between the two formalisms. We choose to do so by reformulating causal models in situation calculus.

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 $(y = C_1 \vee \ldots \vee y = C_n)$. If $X \in \mathcal{U} \cup \mathcal{V}$ with $\mathcal{R}(X) = \{V_1, \ldots, V_n\}$, $y \in \mathcal{R}(X)$ denotes $(y = V_1 \vee \ldots \vee y = V_n)$. 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

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

where $\mathcal{U} \cup \mathcal{V} = \{Z_1, \ldots, Z_N\}$, meaning that the value of variable x is *determined* in s to be v whenever 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:

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

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. Since preconditions disallow value reassignment, the SSA has no negative effects. Overall, the theory models all possible propagations of values throughout the set of variables according to the structural equations, as well as all propagations of values under interventions when some of the variables are forced to specified values. As we are interested only in those situations where all variables have acquired values, which represent a unique solution to $\mathcal{F}$ (possibly after interventions),

we introduce the 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 schema $interv_{Y_1 \leftarrow V_{Y_1}, \ldots, Y_k \leftarrow V_{Y_k}}(s)$ 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 the empty intervention. Notice that in any situation term S that satisfies $interv_{Y_1 \leftarrow V_{Y_1}, \ldots, Y_k \leftarrow V_{Y_k}}(s)$ all actions are executable, but since S is terminal, no further actions are possible, S mentions an action $set(Y, V_Y)$ for every $(Y \leftarrow V_Y)$, and all other actions in S are get.

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 $V(X, V_X, s)$. Notice that $\hat{\phi}$ is ground in all object arguments and uniform in s. This completes the translation. We can now prove the correctness of our axiomatization 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 $\mathcal{D} \models \exists s(interv_{\bar{Y} \leftarrow \bar{V}_Y}(s)) \wedge \forall s(interv_{\bar{Y} \leftarrow \bar{V}_Y}(s) \rightarrow \hat{\phi}(s))$.*

Proof. (Sketch) The proof is straightforward by induction on the structure of ϕ. We prove the base case, when ϕ is a primitive event, by induction on the length of the situation.

By construction, for every $U \in \mathcal{U}$, we have $\mathcal{D} \models V(U, V_U, S_0)$ such that $V_U \in \bar{V}_U$, *i.e.* all exogenous variables have unique, correct values at S_0 (by 'correct' we mean a value which agrees with the causal model). Also by construction, none of the endogenous variables have values at S_0, *i.e.* they are not incorrect.

From this base case, we can always construct an arbitrary narrative σ which conforms to the intervention $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]$ and show that if it does, then it produces only correct values. We make the inductive assumption that there exists a sub-sequence σ' of σ such that all values that exist at σ' are correct. Recall that σ' is not terminal whenever some subset $\mathcal{V}'$ of the endogenous variables have not yet acquired values, *i.e.* $\mathcal{D} \models \bigwedge_{X \in \mathcal{V}'} \neg\exists v\, V(X, v, \sigma')$. Formally, the inductive assumption states that for every $X \in \mathcal{V} \setminus \mathcal{V}'$ and every $V_X \in \mathcal{R}(X)$, $\mathcal{D} \models V(X, V_X, \sigma')$ if and only if $(M, \bar{V}_U) \models [Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k](X = V_X)$. If $\mathcal{V}'$ is empty, then $\sigma' = \sigma$ and our claim follows immediately. Otherwise, we arbitrarily select the next action among the following two options. Option 1: If one of the variables intervened upon does not have a value at σ', *i.e.* $Y_i \in \mathcal{V}'$ for some $1 \leq i \leq k$, then $set(Y_i, V_{Y_i})$ is possible and trivially creates a correct value. Option 2: If there exists some $Z \in \mathcal{V}'$ which is not among $Y_1, \ldots, Y_k$ and $\mathcal{D} \models Det(Z, V_Z, \sigma')$ for some V_Z, then the action $get(Z, V_Z)$ is possible. By the definition of Det, V_Z is obtained from the values which exist at σ' (which we assumed are correct) using the structural equation for Z (which is the same for both BAT and causal model), so the computed value V_Z for Z must agree with the causal model. Since the original causal model has a unique solution, because it is acyclic, if $\sigma' \neq \sigma$, then at least one of these options is always applicable. □

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 $(M, \bar{V}_U)$. 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 $\langle \sigma, \hat{\phi}(s) \rangle$.*

Proof. We sketch the proof of Part 1 only; Part 2 is similar but more involved. As before, we prove the case where ϕ is a primitive event and leave the rest to induction on the structure of ϕ. In the sketch below, for the sake of simplicity, when we talk about single-step regression of formulas that include $Poss(get(X, V_X), s)$ as one of the conjuncts, we omit the terms related to $\neg\exists v' V(X, v', s)$ since they are determined by the initial theory, and keep only $Det(X, V_X, s)$.

($\Rightarrow$.) Suppose $(X = V_X)$ is a singleton cause of ϕ in $(M, \bar{V}_U)$ according to HP^m with a witness $(\bar{W}, \bar{V}_W, V'_X)$. Take an arbitrary ground situation term σ of the BAT $\mathcal{D}$, obtained from $(M, \bar{V}_U)$, such that $\mathcal{D} \models interv_{\emptyset}(\sigma)$. Let $\sigma^\star$ be a terminal ground situation which coincides with σ up to and excluding $get(X, V_X)$, contains $set(X, V'_X)$ in its place, and contains $set(W, V_W)$ in the places of $get(W, V_W)$ for all $W \in \bar{W}$. By the definition of a witness, $(M, \bar{V}_U) \models [X \leftarrow V'_X, \bar{W} \leftarrow \bar{V}_W]\neg\phi$. By Theorem 1, $\mathcal{D} \models \hat{\phi}(\sigma)$ and $\mathcal{D} \models \neg\hat{\phi}(\sigma^\star)$. By construction, all fluent values in $\sigma, \sigma^\star$ agree up to the action $get(X, V_X)/set(X, V'_X)$, so the divergence of the values is accounted for either by V'_X or by some subsequent divergent value. We can show that the difference between V_X and V'_X can explain the divergence of the values. More formally, let ϕ be $Z = V_Z$; recall, we assume that effect ϕ is a primitive event. Since $(M, \bar{V}_U) \models [X \leftarrow V'_X, \bar{W} \leftarrow \bar{V}_W]\neg(Z = V_Z)$, then $(Z = V'_Z)$ must hold under the same intervention for some $V'_Z \neq V_Z$. The primary achievement cause in $\langle \sigma, V(Z, V_Z, s) \rangle$ is the action $get(Z, V_Z)$, which yields a new setting $\langle \sigma', Det(Z, V_Z, s) \rangle$. Since Z acquires a different value in $\sigma^\star$, the achievement cause of the new causal setting must occur in σ no earlier than $get(X, V_X)$. If it is $get(X, V_X)$, we are finished. Otherwise, the same argument applies: we locate the achievement cause, some action $get(Y, V_Y)$ occurring no earlier than $get(X, V_X)$, with $\mathcal{D} \models V(Y, V'_Y, \sigma^\star)$, $V'_Y \neq V_Y$, and generate a new causal setting. Since σ is finite, the analysis converges to the case where the only possible cause is $get(X, V_X)$.

($\Leftarrow$.) For an arbitrary $\sigma = do([get(Z_1, V_{Z_1}), \ldots, get(Z_n, V_{Z_n})], S_0)$ and an arbitrary Z_k $(k \leq n)$ such that $\mathcal{D} \models interv_{\emptyset}(\sigma) \wedge V(Z_k, V_{Z_k}, \sigma)$, the primary achievement cause of $\langle \sigma, V(Z_k, V_{Z_k}, s) \rangle$ is always the action $get(Z_k, V_{Z_k})$ executed in some $\sigma' \sqsubset \sigma$. The remainder of the causal chain is discovered recursively through the new causal setting $\langle \sigma', Det(Z_k, V_{Z_k}, s) \rangle$. The remainder

is empty only if Z_k is determined to be V_{Z_k} from the outset by the context. Otherwise, $\langle \sigma', Det(Z_k, V_{Z_k}, s)\rangle$ has an achievement cause of its own. Since σ represents an empty intervention, this secondary cause is $get(Z_m, V_{Z_m})$ for some $m < k$. Observe that, by the definition of Det, $get(Z_m, V_{Z_m})$ is a cause precisely because the act of setting Z_m to V_{Z_m} removes any F_{Z_k}-borne ambiguity as to the value of Z_k. In other words, prior to $get(Z_m, V_{Z_m})$, the variable Z_k could attain any of at least two possible values, but $get(Z_m, V_{Z_m})$ constrained it to eventually attain V_{Z_k}. Therefore, there exists a (counterfactual) value $V'_{Z_m} \in \mathcal{R}(Z_m)$, $V'_{Z_m} \neq V_{Z_m}$ which, if substituted for V_{Z_m}, would lead to the value of Z_k being different from V_{Z_k}. Let $\sigma^\star$ be an alternative terminal situation which has the same actions as σ up to and excluding $get(Z_m, V_{Z_m})$; the latter is replaced in $\sigma^\star$ by $set(Z_m, V'_{Z_m})$, and all actions $get(Z_j, V_{Z_j})$ for $m < j < k$ are replaced with $set(Z_j, V_{Z_j})$. Notice that the value of the cause is modified, while the subsequent values are forced to stay as they were. Then $\mathcal{D} \models \neg V(Z_k, V_{Z_k}, \sigma^\star)$. Observe that the tuple $((Z_{m+1}, \ldots, Z_{k-1}), (V_{Z_{m+1}}, \ldots, V_{Z_{k-1}}), V'_{Z_m})$ constitutes a *witness* to the fact that $(Z_m = V_{Z_m})$ is an actual cause of $(Z_k = V_{Z_k})$ according to HP^m.

This argument extends by induction to causal chains of arbitrary lengths. Assume that $\langle \sigma'', \psi(s)\rangle$ is a causal setting generated as the result of discovering a sequence of achievement causes $get(Y_1, V_{Y_1}), \ldots, get(Y_t, V_{Y_t})$ via Definition 3 such that the occurrence of $get(Y_j, V_{Y_j})$ precedes the occurrence of $get(Y_{j+1}, V_{Y_{j+1}})$ in σ, *i.e.*, $get(Y_1, V_{Y_1})$ occurs earlier than subsequent causes. The formula $\psi(s)$ is the conjunction $Det(Y_1, V_{Y_1}, s) \wedge \bigwedge_{j=2}^{t} H(j)$ where each conjunct $H(j)$, $j > 1$, is the result of recursively applying single-step regression to $Det(Y_j, V_{Y_j}, s)$ over the actions $get(Y_{j-1}, V_{Y_{j-1}}), \ldots, get(Y_1, V_{Y_1})$ in order, because regression is done over later actions before it is done over earlier actions. This syntactic operation merely replaces the subexpression $V(y, v', s)$ in the definition of $Det(Y_j, V_{Y_j}, s)$ by the expression $\bigvee_{l=1}^{j-1}(y = Y_l \wedge v' = V_{Y_l}) \vee V(y, v', s)$, effectively binding the values of the causal chain discovered so far to the arguments of F_{Y_j}. Now, by the previous argument, if $\langle \sigma'', \psi(s)\rangle$ has a primary achievement cause, then it is some action $get(Y_0, V_{Y_0})$ which eliminates the ambiguity due to the corresponding structural equation in one of the conjuncts of $\psi(s)$. (Note that, by an obvious property of Definition 3, the achievement causal chain of a conjunction contains the achievement causes of all conjuncts.) Therefore, we can exploit the ambiguity and construct an alternative situation where $\psi(s)$ is falsified and extract a HP^m *witness* to this fact. A straight-forward elaboration extends this argument to general (non-atomic) HP queries. □

Example 1 (cont.)**.** Consider a translation of the disjunctive Forest Fire causal model M_d. Recall that neither $(MD = true)$ nor $(L = true)$ alone are singleton actual causes in M_d, but $(MD{=}true) \wedge (L{=}true)$ is an actual cause. Notice that in a previous SC formalization the actions should be renamed to match our translation rules. Namely, replace l with $get(L, true)$, md with $get(MD, true)$, ff with $get(FF, true)$. The corresponding terminal narratives σ are

$$do([get(MD, true), get(L, true), get(FF, true)], S_0),$$
$$do([get(L, true), get(MD, true), get(FF, true)], S_0),$$
$$do([get(MD, true), get(FF, true), get(L, true)], S_0),$$
$$do([get(L, true), get(FF, true), get(MD, true)], S_0).$$

The 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.

6 Discussion

As discussed above, 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 HP's treatment of actual causality (Vennekens, Bruynooghe, and Denecker 2010; Vennekens 2011; Beckers and Vennekens 2012; 2016), but they fail to overcome the expressivity limitations. To our knowledge, the only attempt to lift these limitations was undertaken by (Hopkins 2005; Hopkins and Pearl 2007) who reformulate causal models in the language of situation calculus. In doing so, they arbitrarily designate some causal model variables as 'transitional' and model them as actions, and others as 'enduring' and model them as fluents. In contrast, our translation is systematic and requires no additional modelling decisions. (Hopkins and Pearl 2007) preserve the implicit possible worlds semantics of causal formulas as a layer on top of the (many-sorted version of) standard first order Tarskian semantics of SC and drop the requirement that situations be executable. The latter is especially problematic, since dismissing preconditions results in paradoxes. As an example, consider a BAT modelling the popular Blocks World domain, where the action $move(x, y)$ stacks block x on top of block y and is possible only if there are no blocks on top of x and y; the action $moveToT(x)$ unstacks x and moves it to the table that can hold any number of blocks; the fluent $Clear(x, s)$ states that there is no block on top of block x in situation s; and the fluent $On(x, y, s)$ states that block x is on the top of block y in s. By purging the precondition for $move(x, y)$ from the theory, it is easy to obtain a paradoxical situation $\sigma = do([move(A, B), move(C, B), moveToT(C)], S_0)$ where the theory entails both $Clear(B, \sigma)$ and $On(A, B, \sigma)$. In this case, the query about the presence of something on top of B may yield two opposite answers, depending on how the modeller phrases it. We doubt that one can build a robust definition of actual causality on such shaky foundations. (Hopkins and Pearl 2007) neither attempted to give a formal definition of actual causality, nor provided connections with the causal models approach, as we did.

Curiously, (Vennekens, Bruynooghe, and Denecker 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 statement and

to see where we stand with respect to other approaches, let us consider two telling examples featured in (Beckers and Vennekens 2012). Assume all fluents are false at S_0.

Example 2. *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 before the effect from backup's action can occur. Let the actions be $assassin$ and $backup$ (the two acts of poisoning the coffee) and self-explanatory $drink$. Let the fluents be $P(s)$ meaning "coffee contains poison" and $D(s)$ meaning "the victim is dead".

$$Poss(assassin, s),$$
$$Poss(backup, s),$$
$$Poss(drink, s) \leftrightarrow P(s),$$
$$P(do(a, s)) \leftrightarrow a = assassin \vee a = backup \vee P(s),$$
$$D(do(a, s)) \leftrightarrow [a = drink \wedge P(s)] \vee D(s).$$

The narrative $\sigma = do([assassin, drink], S_0)$ describes the given scenario and $\mathcal{D} \models D(\sigma)$. 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*, and were recognized as such by our analysis. Regardless of how reliable the assassin's backup is, he played no role. The action $assassin$ explains exactly how the victim died; it is an *actual* cause.

Example 3. *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 is proposed by N.Hall (Hall 2000; Paul and Hall 2013) to illustrate the distinction between causation and determination of a causal route; its variants are discussed in many publications (Pearl 2000; Halpern and Pearl 2005; Weslake 2013). Beckers and Vennekens point out that this example is isomorphic to the previous one, except that the intuition about its causes is the polar opposite of that in "Assassin". As we shall see, this dilemma is illusory, and the two examples are isomorphic only within the limited expressivity bounds of causal models and CP-logic. In "Assassin", there are two competing actions, whereas here there is an action and its absence, a distinction which SC is well equipped to capture.

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.

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

The narrative σ is $do([flip, fork_1, fork_2, arrive], S_0)$. The causal setting $\langle \sigma, Arrived(s) \rangle$ has a cause $arrive$. The next setting is $\langle do([flip, fork_1, fork_2], S_0), Out(s) \rangle$ with a cause $fork_2$. Computing $\rho[Out(s), fork_2] \wedge Poss(fork_2, s)$ by Definition 3 yields a new setting $\langle do([flip,\ fork_1],\ S_0),\ L(s) \vee R(s) \rangle$ with a cause $fork_1$. The final setting is $\langle do([flip],\ S_0), \psi(s) \rangle$ where $\psi(s)$ is $(Sw(s) \vee L(s)) \vee (\neg Sw(s) \vee R(s))$ which is a tautology and yields no further causes. Therefore, the $flip$ action is not an actual cause of train's arrival in a faithful SC model of this example, no matter whether the action $flip$ is executed or not. This conclusion is elaboration tolerant (McCarthy 1998) as long as the relation between L, R, Sw is preserved. For HP, the answer depends on how the model is constructed and which definition is applied. (Pearl 2000) calls this class of problems "switching causation" and argues that flipping the switch is a cause (see Section 10.3.4, p.324–5). In a simplified setting with the propositional causal variables $Sw, L, R, Arrived$, consider the equations $Sw := true$, $L := Sw$, $R := \neg Sw$, $Arrived := L \vee R$. According to the original and updated definitions of actual cause, both (Pearl 2000) and (Halpern and Pearl 2005) argue that the switch is a cause. But according to the HP^m definition, the switch is not a cause (Halpern 2016). If we start with this reduced causal model with four variables and translate it to a BAT as proposed in Section 5, we would get the same conclusion as in HP^m, *i.e.*, Theorem 2 applies here too.

The treatment of causality in (Vennekens 2011) is somewhat clouded by using probabilistic rules of the CP-logic, but in fact actual causality can be defined without appeal to probability (Halpern and Pearl 2005; Halpern 2016). (Beckers and Vennekens 2016) introduce concepts of dependence, contribution and production to define basic principles for analysis of actual causality, but their language remains inexpressive with no distinction between properties and actions, and quantified effects are not allowed either.

Example 4. For an example of a quantified query, consider a world with the blocks $\{B_1, B_2, B_3, \ldots\}$ axiomatized so that they form an infinite chain. Let the fluents be $On(x,y,s)$, block x is on block y, $Clear(x,s)$, x is clear, $OnTable(x,s)$, x is on the table. Let the actions be $move(x,y)$, move x on

y, $moveToT(x)$, move x to the table.

$Poss(move(x,y),s) \leftrightarrow Clear(x,s) \wedge Clear(y,s)$,

$Poss(moveToT(x),s) \leftrightarrow Clear(x,s) \wedge \exists y On(x,y,s)$,

$On(x,y,do(a,s)) \leftrightarrow a{=}move(x,y) \ \vee \ On(x,y,s) \wedge$
$\neg\exists z(a{=}move(x,z)) \wedge a{\neq}moveToT(x)$,

$OnTable(x,do(a,s)) \leftrightarrow a{=}moveToT(x) \vee$
$OnTable(x,s) \wedge \neg\exists y\,(a{=}move(x,y)\,)$,

$Clear(x,do(a,s)) \leftrightarrow \exists y,z(a{=}move(y,z){\wedge}On(y,x,s)){\vee}$
$\exists y(a{=}moveToT(y) \wedge On(y,x,s)) \vee$
$Clear(x,s) \wedge \neg\exists y(a{=}move(y,x))$.

Let us assume that initially all blocks are on the table. To show that we can handle quantified causal queries, consider the narrative $\sigma = do([move(B_1,B_2), move(B_1,B_3)], S_0)$ and the effect $\exists x(on(B1,x,s))$. It is easy to see that according to our definition, the first action $move(B_1,B_2)$ is an actual cause of the effect, while the second action is not, since it was preempted by the first action.

In addition to actual achievement causes, it is natural to consider actual maintenance causes. These are the causes responsible for protecting a previously achieved effect, despite potential threats that could destroy the effect. For example, a mitigating action serves as a maintenance cause when it is executed before a threat occurs in a narrative. Our paper (Batusov and Soutchanski 2017) investigates how the notions of achievement and maintenance causes can be combined together into a general definition of an actual cause.

SC includes foundational axioms Σ formulated in second-order logic. However, according to Theorem 1 in (Pirri and Reiter 1999), a BAT $\mathcal{D}$ is satisfiable iff $\mathcal{D}_{S_0} \cup UNA$ is. Additionally, according to Theorem 3 in their paper (Regression Theorem 4.5.5 in (Reiter 2001)), a regressable sentence is entailed by a BAT $\mathcal{D}$ iff the regressed sentence is entailed by $\mathcal{D}_{S_0} \cup UNA$ alone, and if $\mathcal{D}_{S_0}$ is formulated in first order logic (FOL) then this can be reduced to theorem proving in FOL. Moreover, as Reiter argues in Section 4.8 of (Reiter 2001)), in practical applications, when actions have unconditional effects, or when context conditions are situation-free formulas that are decidable wrt $\mathcal{D}_{S_0} \cup UNA$, then the computational complexity of answering projection queries using regression adds at most linear complexity to the complexity of evaluating ground fluents wrt $\mathcal{D}_{S_0} \cup UNA$. The decidability condition is easily satisfied in the case when the object domain is finite, but there are other fragments of SC where the projection problem is reduced to a decidable entailment problem. (Eiter and Lukasiewicz 2002) established that the complexity of deciding whether $X = x$ is a cause of $Y = y$ is NP-complete in Boolean causal models using an older definition of actual cause. More recently, (Aleksandrowicz et al. 2017) explored the complexity of computing actual causes according to the modified definition HP^m.

7 Concluding Remarks

Despite its ingenuity and demonstrated utility, the HP analysis based on causal models has its drawbacks. There exist multiple examples for which 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.

A weakness specific to the HP approach, pointed out by (Glymour et al. 2010) and somewhat mended by (Beckers and Vennekens 2012), stems from its disregard for the order of events which transpire in the given scenario. Such valuable information should not be discarded without a good reason. We believe that this is the essential methodological difference between their approach and ours.

Our work reaps the benefits which (Hopkins and Pearl 2007) aimed at in their choice of situation calculus as the modelling language, but does not suffer from the issues associated with attempting a meaningful definition of a counterfactual in situation calculus, which appears to be no easy task. A counterfactual query not relativized to a particular scenario can be formulated in situation calculus without appealing to 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 situation calculus, but it operates outside of the well-studied basic action theories and is not concerned with actual causality.

The seminal line of enquiry into actual causality stems from Hume and includes such works as (Mackie 1965) and (Lewis 1974). Aside from the aforementioned works, original computational accounts of actual causality are rare, owing, perhaps, to the ubiquity and the appealing simplicity of causal models. 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.

It is clear that a broader 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, *e.g.*, see (Halpern 2016).

Acknowledgements

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

References

Aleksandrowicz, G.; Chockler, H.; Halpern, J. Y.; and Ivrii, A. 2017. The computational complexity of structure-based causality. *J. Artif. Intell. Res.* 58:431–451.

Batusov, V., and Soutchanski, M. 2017. Situation calculus semantics for actual causality. In *13th International Symposium on Commonsense Reasoning*. University College London, UK. Monday, November 6.

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

Beckers, S., and Vennekens, J. 2016. A principled approach to defining actual causation. *Synthese*. DOI 10.1007/s11229-016-1247-1.

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

Costello, T., and McCarthy, J. 1999. Useful counterfactuals. *Electron. Trans. Artif. Intell.* 3(A):51–76.

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

Finzi, A., and Lukasiewicz, T. 2003. Structure-based causes and explanations in the independent choice logic. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, 225–323. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Glymour, C.; Danks, D.; Glymour, B.; Eberhardt, F.; Ramsey, J.; Scheines, R.; Spirtes, P.; Teng, C. M.; and Zhang, J. 2010. Actual causation: a stone soup essay. *Synthese* 175(2):169–192.

Hall, N. 2000. Causation and the price of transitivity. *Journal of Philosophy* 97(4):198–222.

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

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

Halpern, J. Y. 2015. 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*, 3022–3033.

Halpern, J. Y. 2016. *Actual Causality*. The MIT Press, ISBN 9780262035026.

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

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

Hopkins, M. 2005. *The Actual Cause: From Intuition to Automation*. Ph.D. Dissertation, University of California Los Angeles.

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

Lin, F., and Reiter, R. 1994. State constraints revisited. *Journal of logic and computation* 4(5):655–677.

Lin, F., and Soutchanski, M. 2011. Causal theories of actions revisited. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Mackie, J. L. 1965. Causes and conditions. *American Philosophical Quarterly* 2(4):245–264.

McCarthy, J., and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* 4:463–502.

McCarthy, J. 1998. Elaboration tolerance. In *Proc. of the 4th Symposium on Logical Formalizations of Commonsense Reasoning*, 198–216. Queen Mary and Westfield College, University of London, UK. http://www-formal.stanford.edu/jmc/elaboration/.

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

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

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

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1st edition.

Pirri, F., and Reiter, R. 1999. Some contributions to the metatheory of the situation calculus. *Journal of the ACM (JACM)* 46(3):325–361.

Reiter, R. 1991. 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* 359–380.

Reiter, R. 2001. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. The MIT Press, ISBN: 9780262527002.

Simon, H. A. 1953. Causal ordering and identifiability. In Hood, W., and Koopmans, T., eds., *Studies in Econometric Methods*. New York: Wiley. chapter 3, 49–74.

Simon, H. A. 1977. Causal ordering and identifiability. In *Models of Discovery*. Dordrecht: D.Reidel/Springer. 53–80.

Vennekens, J.; Bruynooghe, M.; and Denecker, M. 2010. Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In *European Workshop on Logics in Artificial Intelligence*, 313–325. Springer.

Vennekens, J. 2011. Actual causation in CP-logic. *Theory and Practice of Logic Programming* 11(4-5):647–662.

Weslake, B. 2013. A partial theory of actual causation. In *http://bweslake.s3.amazonaws.com/research/papers/weslake_ac.pdf*. Retrieved on July 18, 2017.