ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200169

Necessary and Sufficient Conditions for Actual Root Causes

Shakil M. Khan and **Mikhail Soutchanski**[1]

Abstract. Reasoning about actual causes of an observed effect is fundamental to many applications. Batusov and Soutchanski (2018) recently presented a first-order logic approach to compute actual causes. Built on a formal theory of action and change, namely the situation calculus, their approach is quite expressive, as it can be used to determine the causes of quantified effects. However, their approach does not find causes from a counterfactual perspective, nor does it link with the regularity approach to causation. This paper proposes a new analysis of actual achievement causes in the situation calculus. We study the natural properties that are necessary for actual causes and conditions that are sufficient for the achievement of an observed (possibly quantified) effect. We identify a property that is both necessary and sufficient for actual achievement causes. This is one of our main contributions. Our discussion leads to a new definition of actual achievement causes that includes the root cause together with a chain of relevant events. We show when our definition is closely related to the recent one proposed by Batusov and Soutchanski (2018).

1 Introduction

Research on actual causality involves finding in a given trace (a log, a record) actions that caused an effect. Pearl [33, 34] was a pioneer to lead a computational enquiry into actual causality. The research was later continued by Halpern and Pearl [14, 18] and others [7, 20, 22, 15, 16]. This "HP approach" is based on the concept of structural equations [36]. HP follows the Humean counterfactual definition of causation, which states that saying that "an outcome B is caused by an event A" is the same as saying that "had A never occurred, B never had existed". This definition suffers from the problem of preemption[2]: 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 trace 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. The approach based on Structural Equations Models (SEM) has been criticized for its limited expressiveness [20, 22, 10], and researchers have attempted to expand SEM with additional features, e.g. [25].

A different approach was recently proposed by Batusov and Soutchanski [1], who developed a foundational definition of actual achievement cause within situation calculus basic action theories [35]. While they focused on linear traces only, an advantage of their approach is that it is based on an expressive formal theory of action and change. Also, it allows one to reason about actual causes of *quantified* effects. However, they neither define an actual cause in counterfactual terms, nor do they relate their definition to the regularity account of causation. As a consequence, it is not clear how their definition can be related to these two common, but different approaches to actual causes. There is strong experimental evidence that humans understand causes using counterfactual reasoning [9, 8]. In contrast, others [6, 2] identified limitations of counterfactual-based reasoning and argued for regularity definitions.

In this paper, we explore a set of necessary properties for actual achievement causes, and discuss sufficient conditions for the achievement of observed effects within the situation calculus. This is one of our contributions. We then identify a property that is both necessary and sufficient for actual achievement causes. This is our main contribution. Using this, we give a new definition of actual achievement causes (Definition 12) that has a counterfactual flavour. We make two simplifying assumptions. First, we deal with achievement causes exclusively. Second, we consider only the sequential case, when the actions are completely ordered. We prove that the identified conditions have some intuitively desirable properties, and investigate the formal relation between our new definition and that of [1]. This is our final contribution. Interestingly, our new definition of actual cause helps illustrate Mackie's [27] well-known INUS condition interpretation of Hume's regularity definition. We elaborate it by performing a temporal check while choosing the appropriate INUS condition, eliminating the problems with cases of preemption.

We start with a motivating example in §2. In §3, we formalize the necessary properties and the sufficient conditions. In §4, we identify the necessary and sufficient property for causes and propose our new definition. We investigate the formal relation between our definition and [1] in §5. In §6 and 7, we discuss related work and conclude.

2 A Motivating Example

The Situation Calculus. The situation calculus (SC) [29] is a popular formalism for modeling and reasoning about dynamic systems. We use a many-sorted version that defines a basic action theory (BAT) $\mathcal{D}$ [35]. We will explain the main ingredients of a BAT using the following motivational example. We will use the complex situation term $do([\alpha_1, \cdots, \alpha_n], S_0)$ to represent the situation obtained by consecutively performing $\alpha_1, \cdots, \alpha_n$ starting from S_0. Also, the notation $s \sqsubset s'$ means that situation s' can be reached from situation s by executing a sequence of actions. $s \sqsubseteq s'$ is an abbreviation of $s \sqsubset s' \vee s = s'$. $s < s'$ is an abbreviation of $s \sqsubset s' \wedge executable(s')$, where $executable(s)$ is defined as $\forall a', s'.\ do(a', s') \sqsubseteq s \rightarrow Poss(a', s')$. We will also utilize a *single-step regression* operator $\rho[\phi, \alpha]$. Given a query "does ϕ hold

[1] Ryerson University, Canada, email: {shakilmkhan,mes}@scs.ryerson.ca
[2] Preemption happens when two competing events try to achieve the same effect, and the latter of these fails to do so, as the earlier one has already achieved the effect.

in the situation obtained by performing the ground[3] action α in some situation σ, i.e. in $do(\alpha, \sigma)$?", ρ transforms it into an equivalent query "does ψ hold in situation σ?", eliminating action α by compiling its effects into ψ, that provides the weakest preconditions of ϕ in σ. In the sequel, we use lowercase Greek letters and uppercase Latin letters for ground terms, and lowercase Latin letters for variables. All free variables in a sentence are assumed to be implicitly $\forall$-quantified.

Example. We use a simple connected car as an example. There is a car C (this is a constant). A driver can drive any car c from intersection i to j by executing the $drive(c, i, j)$ action. The geometry of the roads is captured using the non-fluent relation $conn(i, j)$, which states that there is a street from intersection i to j. To respond to cyber-threats and newly discovered vulnerabilities, each car has the ability to wirelessly download security patches. Unfortunately, due to poor design choices, the cars are still susceptible to cyber-attacks.[4] In particular, there are two hackers in the domain, H_1 and H_2. Hacker H_1, who is located at intersection I_2, is equipped with the capabilities of intercepting a car's key-fob signal, infiltrating its software, and remotely controlling the car, if the car is within H_1's range (i.e. both are at the same intersection). This can be done using the $keyHack(h, c)$ action. Once hacked, the hackers can also erase all security patches by executing the $eraseP(c)$ action, which makes the car vulnerable to over-the-air attacks. On the other hand, hacker H_2, who is driving around, can exploit a bug in (the original version of) a car's on-board software system to intercept its telematics connection and take over the car's Engine Control Modules, provided the software is not up-to-date. Such attacks can be attempted using the $teleHack(h, c)$ action. A hacked car may be reclaimed using the $recover(c)$ action and the latest security patches can be installed using the $installP(c)$ action. Initially, all the cars are vulnerable, but are not hacked.

There are three fluents, $at(c, i, s)$, $vulnerable(c, s)$, and $hacked(c, s)$, which mean that the car c is at location i in situation s, c's on-board software is not up-to-date and as such it is vulnerable to hacks in s, and c is hacked in s, respectively.

We provide axioms for a BAT for this example. First, the action precondition axioms for the aforementioned actions are as follows.

(1) $Poss(drive(c, i, j), s) \leftrightarrow at(c, i, s) \wedge i \neq j \wedge conn(i, j)$,
(2) $Poss(keyHack(h, c), s) \leftrightarrow h = H_1 \wedge \exists i(at(h, i, s) \wedge at(c, i, s))$,
(3) $Poss(teleHack(h, c), s) \leftrightarrow h = H_2 \wedge vulnerable(c, s)$,
(4) $Poss(eraseP(c), s) \leftrightarrow hacked(c, s)$,
(5) $Poss(installP(c), s) \leftrightarrow \neg hacked(c, s)$,
(6) $Poss(recover(c), s) \leftrightarrow hacked(c, s)$.

These are self-explanatory; e.g., (1) states that c can be driven from i to j in situation s if and only if c is at i in s, i and j refer to different intersections, and there is a street connecting i and j.

Moreover, the following successor-state axioms (SSA) specify when exactly the fluents at, $vulnerable$, and $hacked$ change value when an action a is executed in some situation s.

(7) $at(c, i, do(a, s)) \leftrightarrow (\exists j(a = drive(c, j, i)) \vee (at(c, i, s) \wedge \neg\exists j(a = drive(c, i, j))))$,
(8) $vulnerable(c, do(a, s)) \leftrightarrow (a = eraseP(c) \vee (vulnerable(c, s) \wedge a \neq installP(c)))$,
(9) $hacked(c, do(a, s)) \leftrightarrow (\exists h(a = keyHack(h, c) \vee a = teleHack(h, c)) \vee (hacked(c, s) \wedge a \neq recover(c)))$.

That is, (7) states that c is at location i in the situation resulting from executing some action a in situation s (i.e. in $do(a, s)$) if and only if a refers to c's action of driving from location j to i, or it was already at i in s and a is not the action of driving c to another location j. The axioms for $vulnerable$ in (8) and $hacked$ in (9) are similar.

Furthermore, the initial situation is specified using the following initial state axioms. (10) $at(C, I_1, S_0)$, (11) $at(H_1, I_2, S_0)$, (12) $\forall c(vulnerable(c, S_0))$, (13) $\forall c(\neg hacked(c, S_0))$. Thus, e.g. initially the car C is at intersection I_1, etc.

We assume that the intersections I_1 and I_2 are connected. (14) $\forall i, j.\ conn(i, j) \leftrightarrow ((i = I_1 \wedge j = I_2) \vee (i = I_2 \wedge j = I_1))$. Also, for simplicity and illustration, we assume the domain closure axiom for the intersections, stating that there are only two intersections I_1 and I_2 in this domain. (15) $\forall i(i = I_1 \vee i = I_2)$. However, we do not require a domain closure axiom for cars and hackers, as their number can be unknown. We need unique names axioms stating that I_1 and I_2 refer to two different intersections, and similarly for hackers: (16) $I_1 \neq I_2 \wedge H_1 \neq H_2$. Finally, we need unique-names axioms (UNA) stating that $drive$, $keyHack$, $teleHack$, $installP$, $eraseP$, and $recover$ refer to different actions, and that two actions with the same function symbol refer to the same action if their arguments are the same. We omit these for brevity. Henceforth, we use $\mathcal{D}_{cc}$ to refer to the above axiomatization of the car domain.

For an example of single-step regression, let us compute $\rho[\exists c(hacked(c, do(keyHack(H_1, C), S^*)),\ keyHack(H_1, C)]$ for some situation S^*. We substitute action variable a by $keyHack(H_1, C)$ in the right hand side of the successor-state axiom (9), and replace the situation variable s by S^*. This yields $\exists h(keyHack(H_1, C) = keyHack(h, c) \vee keyHack(H_1, C) = teleHack(h, c)) \vee [hacked(c, S^*) \wedge keyHack(H_1, C) \neq recover(c)]$. Using the unique names axioms for actions, this is equivalent to $\exists h(H_1 = h \wedge C = c) \vee hacked(c, S^*)$. So, the query is equivalent to $\exists c \exists h(H_1 = h \wedge C = c \vee hacked(c, S^*))$, and this can be simplified to $true$.

3 Necessary Properties, Sufficient Conditions

Given a trace σ, *actual achievement causes* are actions that are behind achieving some effect. In this section, we propose a set of necessary properties of actual achievement causes and a sufficient condition for actual achievement causality within the SC. An effect in this framework is a SC formula $\phi(s)$ that is uniform in s (meaning that it has no occurrences of $Poss$, $\sqsubset$, other situation terms besides s, and quantifiers over situations). Recall $\mathcal{D}$ denotes a BAT.

Definition 1 (Causal Setting) *A causal setting is a tuple $\langle \mathcal{D}, \sigma, \phi(s) \rangle$, where $\mathcal{D}$ is a theory, σ is a ground situation term of the form $do([\alpha_1, \cdots, \alpha_n], S_0)$ with ground action functions $\alpha_1, \cdots, \alpha_n$ such that $\mathcal{D} \models executable(\sigma)$, and $\phi(s)$ is a SC formula uniform in s such that $\mathcal{D} \models \phi(\sigma)$.*

As the theory $\mathcal{D}$ is fixed, we will often suppress $\mathcal{D}$. Also, here we require ϕ to hold by the end of the trace σ.

Since all changes in the SC result from actions, we identify the potential causes with a set of ground action terms occurring in σ. However, since σ might include multiple occurrences of the same action, we also need to identify the situations where these actions were executed. Thus, a *cause* with respect to a causal setting is a non-empty set of (action, situation) pairs derived from the trace σ. We call each pair in a cause, a *part of the cause*.

For example, consider the trace $\sigma_{cc} = do([teleHack(H_2, C), recover(C), installP(C), drive(C, I_1, I_2), keyHack(H_1, C), eraseP(C), teleHack(H_2, C)], S_0)$. We are interested in comput-

[3] A ground term is one whose constituents are ground sub-terms and constants, i.e. a term that contains no variables.
[4] Most of the following scenarios are realistic; indeed remote car hacking remains a very real threat today [11].

ing the actual causes of the effect $\phi_{cc} = \exists c(hacked(c, s))$. Thus, the causal setting is $\mathcal{C}_1 = \langle \mathcal{D}_{cc}, \sigma_{cc}, \phi_{cc} \rangle$. It is easy to see that the cause of ϕ_{cc} is $\{(drive(C, I_1, I_2), S_3), (keyHack(H_1, C), S_4)\}$, where $S_3 = do([teleHack(H_2,C), recover(C), installP(C)], S_0)$ and $S_4 = do(drive(C, I_1, I_2), S_3)$. Note that the first $teleHack$ action is not a cause, as its effect on ϕ_{cc} did not persist till the end of the trace. The second $teleHack$ is also not a cause since it was preempted by the $keyHack$ action. Finally, the $drive$ action is part of the cause since it is required for the $keyHack$ to work: H_1 could not have hacked the car if the car were not at intersection I_2.

Necessary Properties. We start by informally analyzing a set of properties that an actual achievement cause must necessarily have. Intuition suggests that there are three such necessary properties that a (part of a) cause must have. (N_1) Each (part of a) cause must contribute to the achievement of the effect ϕ. (N_2) A cause must not be preempted w.r.t the effect given trace σ. (N_3) The effect ϕ brought about by a cause must be enduring, i.e. it cannot be the case that a subsequent action on the trace σ makes ϕ false.

We now gradually formalize the necessary properties for actual achievement causes. First, we give a trace-independent definition of contributing causes, both direct and indirect.

Definition 2 (Direct Possible Contributor). *Given theory $\mathcal{D}$ and effect $\phi(s)$, an action α is called a direct possible contributor to $\phi(s)$ if and only if there is a situation σ such that*

$$\mathcal{D} \models executable(\sigma) \wedge Poss(\alpha, \sigma) \wedge \neg\phi(\sigma) \wedge \phi(do(\alpha, \sigma)).$$

The situation σ is called a witness for the contribution of action α relative to theory $\mathcal{D}$ and effect $\phi(s)$.

For example, given domain $\mathcal{D}_{cc}$, it can be shown that two direct possible contributors to $\phi_{cc} = \exists c(hacked(c, s))$ are $keyHack(H_1, C)$ and $teleHack(H_2, C)$, since by Axiom (9), these actions directly contribute to the achievement of the fluent $hacked$. For the former, the situation $do(drive(C, I_1, I_2), S_3)$ is a witness, while for the latter, the situation S_0 is a witness.

Definition 3 (Possible Contributor and Contributing Cause). *Given theory $\mathcal{D}$ and effect $\phi(s)$, an action α_1 is called a possible contributor to $\phi(s)$ if and only if there are non-empty finite sequences of actions $\alpha_1, \cdots, \alpha_n$, situations $\sigma_1, \cdots, \sigma_n$, and formulae $\phi_1, \cdots, \phi_n$ such that **(A)** α_1 is a direct possible contributor to ϕ_1 with witness $\sigma_1, \cdots, \alpha_n$ is a direct possible contributor to ϕ_n with witness σ_n, and **(B)** $\mathcal{D} \models \sigma_1 < do(\alpha_1, \sigma_1) \leq \sigma_2 < do(\alpha_2, \sigma_2) \leq \cdots \leq \sigma_n$, and **(C)** $\mathcal{D} \models \forall s.\ \phi_n(s) \leftrightarrow \phi(s)$, $\mathcal{D} \models \forall s.\ \phi_{n-1}(s) \leftrightarrow Poss(\alpha_n, \sigma_n) \wedge \rho[\phi_n, \alpha_n], \cdots, \mathcal{D} \models \forall s.\ \phi_1(s) \leftrightarrow Poss(\alpha_2, \sigma_2) \wedge \rho[\phi_2, \alpha_2]$, and **(D)** $\mathcal{D} \models \forall s.\ (do(\alpha_1, \sigma_1) \leq s \leq \sigma_2 \rightarrow \phi_1(s)), \cdots, \mathcal{D} \models \forall s.\ (do(\alpha_{n-1}, \sigma_{n-1}) \leq s \leq \sigma_n \rightarrow \phi_{n-1}(s))$. We call the sequence of situations $\{\sigma_1, \cdots, \sigma_n, do(\alpha_n, \sigma_n)\}$ a witness for α_1's contribution.*

Moreover, if $\alpha_1, \cdots, \alpha_n$ is maximal in the sense that there does not exist an action α^*, situation σ^*, and formula ϕ^* such that α^* is a direct contributor to ϕ^* with witness σ^*, where $\mathcal{D} \models \sigma^* < do(\alpha^*, \sigma^*) \leq \sigma_1$ and $\mathcal{D} \models \forall s.\ \phi^*(s) \leftrightarrow Poss(\alpha_1, \sigma_1) \wedge \rho[\phi_1, \alpha_1]$ and $\mathcal{D} \models \forall s.\ do(\alpha^*, \sigma^*) \leq s \leq \sigma_1 \rightarrow \phi^*(s)$, then in addition we call the ordered set $\{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ a possible contributing cause of $\phi(s)$.

Definition 3 formalizes a chain of direct possible contributors (condition *(A)*), where the effect of the final element of the chain is ϕ (the first item in above condition *(C)*) and the first direct contributor is the action α_1 executed in situation σ_1. Thus α_1 indirectly possibly contributes to the effect ϕ. Condition *(B)* specifies how the witnessing situations in this chain are related. Condition *(C)* on the other hand specifies the intermediate effects; these include appropriately regressed effects and preconditions of actions in the chain. Note that, in addition to the contributing actions $\alpha_1, \cdots, \alpha_n$, the trace defined by $(\sigma_1, \ldots, do(\alpha_n, \sigma_n))$ may include other actions that are irrelevant to the contribution to ϕ. Thus, we need to ensure that the intermediate effects are not perturbed by these. This is formalized in condition *(D)* above, which requires that an intermediate effect brought about by an action in the chain persists until the situation where the next relevant action in the chain is executed.

In our example, but without regard to σ_{cc}, a possible indirect contributor to ϕ_{cc} is the action $drive(C, I_1, I_2)$. A witness to this can be $\{S_0, S^*, do(keyHack(H_1, C), S^*)\}$, where $S^* = do([drive(C, I_1, I_2), installP(C)], S_0)$.[5] Note the irrelevant action $installP(C)$ here. This is the case since the conditions in Definition 3 are satisfied by the finite sequence of actions $drive(C, I_1, I_2), keyHack(H_1, C)$. In particular, $keyHack(H_1, C)$ is indeed a direct contributor to ϕ_{cc} with witness S^*. To see this, note that the initial situation S_0 is an executable situation (this is a property of the SC). Moreover, by Axioms (1), (10), (16), and (14), the $drive(C, I_1, I_2)$ action is possible in S_0, and therefore $do(drive(C, I_1, I_2), S_0)$ is also an executable situation. Since initially C is at location I_1 (Axiom (10)), by Axiom (7) and the unique-names axioms, C will be at location I_2 in $do(drive(C, I_1, I_2), S_0)$. Similarly, it can be shown that $installP(C)$ is executable in $do(drive(C, I_1, I_2), S_0)$ and after its execution, C will still be at location I_2. Also, using Axiom (11) and a similar analysis, it can be shown that hacker H_1 will also be at I_2 in S^*. Finally, by these and Axiom (2), the $keyHack(H_1, C)$ action is executable in S^*, and by Axiom (9), it brings about ϕ_{cc}.

Similarly, it can be shown $drive(C, I_1, I_2)$ is a direct contributor to $Poss(keyHack(H_1,C), S^*) \wedge \rho[\phi_{cc}, keyHack(H_1,C)]$, which can be simplified to $\exists i(at(H_1,i,S^*) \wedge at(C,i,S^*))$ with witness S_0.

Moreover, it can be shown that all other conditions in Definition 3 are also fulfilled. Thus, $drive(C, I_1, I_2)$ with witness $\{S_0, S^*, do(keyHack(H_1, C), S^*)\}$ is indeed a possible contributor to ϕ_{cc}. Note that the trace defined by $(S_0, \ldots, do(keyHack(H_1, C), S^*))$ cannot be extended in the past, since there is no situations before S_0. Thus, $\{(drive(C, I_1, I_2), S_0), (keyHack(H_1, C), S^*)\}$ is a possible contributing cause of ϕ_{cc}.

We next give definitions of direct and indirect contributing causes that take the given narrative into account.

Definition 4 (Direct Actual Contributor). *Given a causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$, an action α^* is called a direct actual contributor to $\phi(s)$ if and only if α^* is a direct possible contributor to $\phi(s)$ given $\mathcal{D}$, σ^* is a witness to this, and the following holds.*

$$\mathcal{D} \models S_0 \leq \sigma^* < do(\alpha^*, \sigma^*) \leq \sigma.$$

The situation σ^* is called a witness for action α^*'s direct actual contribution w.r.t $\mathcal{C}$.

Note that the condition in the above definition means that α^* actually occurs in the trace σ.

[5] This example is not related to the trace σ_{cc} since here we are talking about *possible* contributors, but not about actual contributors.

For example, it can be shown that there are only two direct actual contributors to ϕ_{cc} in the trace σ_{cc}, namely $teleHack(H_2, C)$ with witness S_0 and $keyHack(H_1, C)$ with witness S_4. More importantly, $teleHack(H_2, C)$ executed in $S_6 = do([keyHack(H_1, C), eraseP(C)], S_4)$ is not a direct actual contributor. To see this notice that a direct actual contributor is also a direct possible contributor, and therefore it must trigger the truth value of the effect from $false$ to $true$. Using arguments as before, it can be shown that ϕ_{cc} was true in S_6. Hence, the $teleHack$ action executed in S_6 cannot be a direct actual contributor.

Definition 5 (Actual Contributor and Actual Contributing Cause). *Given a causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$, an action α_1 is called an actual contributor to $\phi(s)$ if and only if α_1 is a possible contributor to $\phi(s)$ given $\mathcal{D}$, $\{\sigma_1, \cdots, \sigma_n, do(\alpha_n, \sigma_n)\}$ is the witness to this, and the following holds:*

$$\mathcal{D} \models S_0 \leq \sigma_1 \leq \cdots \leq \sigma_n \leq do(\alpha_n, \sigma_n) \leq \sigma.$$

We call the sequence $\{\sigma_1, \cdots, \sigma_n\}$ a witness for α_1's contribution.

Moreover, similar to Definition 3, we also call the action-situation pairs $\{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ an actual contributing cause when the sequence $\alpha_1, \cdots, \alpha_n$ is maximal.

This notion of actual contributing cause captures necessary properties N_1 and N_2 discussed above. By definition, each part of an actual contributing cause contributes to the effect ϕ. Moreover, an actual contributing cause cannot be preempted, since it must include an action –the final one in the chain – that directly actually (and thus possibly) contributes to ϕ, i.e. changes the truth value of ϕ from false to true (see Definition 2). Consequently, if ϕ was previously achieved in σ by another contributor, its contribution was not enduring.

For example, in the trace σ_{cc}, $\{(teleHack(H_2, C), S_0)\}$ and $\{(drive(C, I_1, I_2), S_3), (keyHack(H_1, C), S_4)\}$ are the only two actual contributing causes. Once again, the second $teleHack$ action executed in S_6 cannot be included as a part of an actual contributing cause, as the effect ϕ_{cc} was already $true$ in S_6.

We refine this notion to include the necessary property N_3.

Definition 6 (Direct Enduring Producer). *Given a causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$, an action α^* is called a direct enduring producer of $\phi(s)$ if and only if α^* is a direct actual contributor to $\phi(s)$ given $\mathcal{C}$, σ^* is the witness to this, and*

$$\mathcal{D} \models \forall s.\ \sigma^* < s \leq \sigma \rightarrow \phi(s).$$

The situation σ^* is called a witness for action α^*'s direct enduring production w.r.t $\mathcal{C}$.

We can show that while $keyHack(H_1, C)$ executed in S_4 is a direct enduring producer, $teleHack(H_2, C)$ executed in S_0 is not. This is because the effect ϕ_{cc} brought about by the first $teleHack$ action is reversed by the action $recover(C)$ executed in $S_1 = do(teleHack(H_2, C), S_0)$, and as such its effects are not enduring.

Definition 7 (Enduring Producer). *Given a causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$, an ordered set of action-situation pairs $\{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ is called an enduring producer of $\phi(s)$ if and only if $\{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ is an actual contributing cause of $\phi(s)$ given $\mathcal{C}$, $\{\sigma_1, \cdots, \sigma_n\}$ is the witness to this, and the following holds.*

$$\mathcal{D} \models \forall s.\ \sigma_n < s \leq \sigma \rightarrow \phi(s).$$

We call the situation sequence $\{\sigma_1, \cdots, \sigma_n\}$ a witness for $\{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$'s enduring production.

We can show that the only enduring producer in our example is $\{(drive(C, I_1, I_2), S_3), (keyHack(H_1, C), S_4)\}$.

We now formally show that all necessary properties of actual achievement causes suggested above hold for enduring producers. We start with necessary property N_1. If $\mathcal{K} = \{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ is an ordered set of action-situation pairs, let $\vec{\alpha}_{\mathcal{K}}$ refer to the sequence of actions $(\alpha_1, \cdots, \alpha_n)$. Let $\vec{\alpha}_{sub\{\mathcal{K}\}}$ denote any proper subsequence of $\vec{\alpha}_{\mathcal{K}}$ that does not alter the order of the actions in $\vec{\alpha}_{\mathcal{K}}$.

Proposition 1 (Contribution of Enduring Producers). *Let $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$ be a causal setting, and $\mathcal{K} = \{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ be an enduring producer of $\mathcal{C}$. Then*

$$\mathcal{D} \models \big(executable(do(\vec{\alpha}_{\mathcal{K}}, S_0)) \wedge \phi(do(\vec{\alpha}_{\mathcal{K}}, S_0)) \big) \wedge \neg \big(executable(do(\vec{\alpha}_{sub\{\mathcal{K}\}}, S_0)) \wedge \phi(do(\vec{\alpha}_{sub\{\mathcal{K}\}}, S_0)) \big).$$

Thus, every action in an enduring producer $\mathcal{K}$ directly or indirectly contributes to ϕ in that no proper subsequence of $\vec{\alpha}_{\mathcal{K}}$ is sufficient to bring about ϕ. Next, let us consider necessary property N_2.

Proposition 2 (Uniqueness of Enduring Producers). *Let $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$ be a causal setting, and $\mathcal{K}_1$ and $\mathcal{K}_2$ be two enduring producers of $\mathcal{C}$. Then $\mathcal{K}_1 = \mathcal{K}_2$.*

Thus, since an enduring producer cannot be preceded by another enduring producer, any actual contributing cause that may come before an enduring producer must be non-persistent. So, an enduring producer cannot be preempted.

Finally, observe that necessary property N_3 trivially follows from the definition of enduring producers.

Proposition 3 (Persistence of Enduring Producer Effects). *Let $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s) \rangle$ be a causal setting, and $\mathcal{K} = \{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ be an enduring producer of $\mathcal{C}$. Then $\mathcal{D} \models \forall s.\ \sigma_n < s \leq \sigma \rightarrow \phi(s)$.*

Sufficient Conditions. Let us now turn our attention to conditions that are sufficient for actual achievement causation. A reasonable condition for an ordered set of (action, situation) pairs to cause an effect ϕ is that given trace σ, the execution of these actions must bring about ϕ, and the achieved effect ϕ disappears if any of the actions on which ϕ depends are withdrawn.

In the following, we formalize these conditions. If $\sigma = do([\alpha_0, \alpha_1, \cdots, \alpha_n], S_0)$ is a ground situation, then let $\langle \sigma \rangle$ be the ordered set $\{(\alpha_0, S_0), (\alpha_1, S_1), \cdots, (\alpha_n, S_n)\}$, where $S_1 = do(\alpha_0, S_0)$, etc. Also, let $\langle \sigma_{\overline{(\alpha^*, \sigma^*)}} \rangle = \langle \sigma \rangle \setminus \{(\alpha^*, \sigma^*)\}$, where $(\alpha^*, \sigma^*) \in \langle \sigma \rangle$, and $\sigma_{\overline{(\alpha^*, \sigma^*)}}$ be the associated situation. Informally, we use the horizontal bar over a subscript of an ordered set to denote the set-theoretic difference when the subscript is removed from the set. Thus, $\sigma_{\overline{(\alpha^*, \sigma^*)}}$ is the situation that can be obtained by "executing" all the actions in σ except for α^* in situation σ^* in the order they appear in σ, starting in S_0. For instance, if $\sigma = do([\alpha, \beta, \gamma, \delta, \beta, \epsilon], S_0)$, then $\sigma_{\overline{(\beta, S_1)}} = do([\alpha, \gamma, \delta, \beta, \epsilon], S_0)$, where $S_1 = do(\alpha, S_0)$. We start with a tentative naïve version.

Definition 8 (Counterfactual Dependence (Naïve Version)). *Given setting $\langle \mathcal{D}, \sigma, \phi(s) \rangle$, action α^*, and situation σ^*, such that*

$$\mathcal{D} \models executable(\sigma) \wedge \phi(\sigma) \wedge S_0 < do(\alpha^*, \sigma^*) \leq \sigma,$$

ϕ is counterfactually dependent on α^* executed in σ^* if and only if

$$\mathcal{D} \models \neg executable(\sigma_{\overline{(\alpha^*,\sigma^*)}}) \vee \neg\phi(\sigma_{\overline{(\alpha^*,\sigma^*)}}).$$

Note that this definition requires us to take into account the executability of actions. The above tentative definition is not good enough, since the non-executability of $\sigma_{\overline{(\alpha^*,\sigma^*)}}$ does not necessarily imply counterfactual dependence. For instance, it can be the case that α^* executed in σ^* is totally irrelevant w.r.t the achievement of ϕ, but it makes the precondition of another irrelevant action β false. Thus, we need to ensure that such cases are accounted for.

Let $\langle\sigma_{\overline{Cl(\alpha^*,\sigma^*)}}\rangle$ denote the ordered set $\langle\sigma\rangle \setminus \langle Cl(\alpha^*,\sigma^*,\sigma)\rangle$, where $\langle Cl(\alpha^*,\sigma^*,\sigma)\rangle$ is the least set $\mathcal{P}$ such that (α^*,σ^*) is in $\mathcal{P}$, and if (α',σ') is in $\mathcal{P}$ and there exists an action α'' and a situation σ'' such that $\mathcal{D} \models S_0 \sqsubseteq \sigma'' \sqsubset do(\alpha'',\sigma'') \sqsubseteq \sigma_{\overline{(\alpha',\sigma')}} \wedge \neg Poss(\alpha'',\sigma'')$, then (α'',σ'') is also in $\mathcal{P}$.[6] Let $\sigma_{\overline{Cl(\alpha^*,\sigma^*)}}$ be the associated situation. Thus, $\sigma_{\overline{Cl(\alpha^*,\sigma^*)}}$ is the situation obtained by executing all the actions in σ starting in S_0 in the order they appear in σ, except for action α^* executed in situation σ^*, and except for all subsequent actions whose preconditions are broken by the removal of α^* in σ^* from σ. For example, if $\sigma = do([\alpha,\beta,\gamma,\delta,\beta,\epsilon],S_0)$ and the removal of β executed in $S_1 = do(\alpha,S_0)$ from σ only makes the preconditions of δ false, then $\sigma_{\overline{Cl(\beta,S_1)}} = do([\alpha,\gamma,\beta,\epsilon],S_0)$.

In the following, we propose an improved definition.

Definition 9 (Counterfactual Dependence). *Given causal setting $\langle\mathcal{D},\sigma,\phi(s)\rangle$, action α^*, and situation σ^*, such that*

$$\mathcal{D} \models executable(\sigma) \wedge \phi(\sigma) \wedge S_0 < do(\alpha^*,\sigma^*) \leq \sigma,$$

ϕ is counterfactually dependent on α^* executed in σ^* if and only if

$$\mathcal{D} \models \neg\phi(\sigma_{\overline{Cl(\alpha^*,\sigma^*)}}).$$

We call situation σ^* the witness for ϕ's dependence on α^*.

The above definition specifies that ϕ counterfactually depends on α^* executed in σ^* if and only if removing α^* along with all other actions whose preconditions (directly or indirectly) depend on α^* in σ^* yields a situation where ϕ does not hold. Thus, ϕ must have been directly or indirectly dependent on α^* in σ^*.

In our example, ϕ_{cc} counterfactually depends on the action $drive(C,I_1,I_2)$ in σ_{cc}. To see this, note that removing $drive(C,I_1,I_2)$ from the trace σ_{cc} along with other actions that no longer will be executable, i.e., $keyHack(H_1,C)$, $eraseP(C)$, and the second $teleHack(H_2,C)$, yields the new executable trace $[teleHack(H_2,C), recover(C), installP(C)]$. It can be shown that ϕ_{cc} is false after executing these actions from S_0.

Clearly, this notion of counterfactual dependence is not good enough for causation (this is not to say that the above definition of counterfactual dependence is problematic). While counterfactual dependence of an effect ϕ on an action α^* executed in some situation σ^* ensures that α^* executed in σ^* is a (part of a) cause of ϕ, it does not entail that α^* in σ^* *alone* is guaranteed to provide for ϕ, as we may need additional actions besides α^* to bring about ϕ; e.g. $drive(C,I_1,I_2)$ by itself is not sufficient for ϕ_{cc} as $keyHack(H_1,C)$ is also needed for the effect.

Hence, we need to enhance the previous sufficiency condition.

Definition 10 (Counterfactual Dependence w.r.t Actual Contributing Cause). *Given setting $\mathcal{C} = \langle\mathcal{D},\sigma,\phi(s)\rangle$ and an actual contributing cause $\mathcal{K} = \{(\alpha_1,\sigma_1),\cdots,(\alpha_n,\sigma_n)\}$ of $\mathcal{C}$, ϕ is counterfactually dependent on $\mathcal{K}$ if and only if*

$$\mathcal{D} \models \neg\phi(\sigma_{\overline{Cl(\alpha_n,\sigma_n)}}).$$

That is, ϕ counterfactually depends on an actual contributing cause $\mathcal{K}$ if and only if removing the last action of $\mathcal{K}$ along with all other actions whose preconditions depend on it yields a situation where ϕ does not hold. Note that the final action in the chain $\mathcal{K}$ is indeed the one that is ultimately responsible for achieving the effect ϕ.

Obviously, ϕ_{cc} counterfactually depends on the actual contributing cause $\{(drive(C,I_1,I_2),S_3),\ (keyHack(H_1,C),S_4)\}$ since removing the final action, i.e. $keyHack(H_1,C)$, along with others that depend on this action, from the original trace yields the new trace $[teleHack(H_2,C), recover(C), installP(C), drive(C,I_1,I_2)]$, whose execution starting from S_0 does not bring about ϕ_{cc}. On the other hand, ϕ_{cc} does not counterfactually depend on the actual contributing cause $\{(teleHack(H_2,C),S_0)\}$, since removing this action from the original trace gives us a trace whose execution still brings about ϕ_{cc}; this new trace is simply all actions in σ_{cc} except for the first two actions, $teleHack$ and $recover$.

Note that, counterfactual dependence w.r.t actual contributing cause is a sufficient but unnecessary condition for causation. It is sufficient in the sense that the presence of such an actual contributing cause on the trace σ (on which ϕ is dependent) guarantees ϕ. It is unnecessary since there are cases where an effect ϕ is not counterfactually dependent on some actual contributing cause $\mathcal{K}$, but $\mathcal{K}$ is a cause of ϕ nonetheless, e.g. when there is another independent and competing action, whose effect on ϕ is preempted by $\mathcal{K}$ in σ, and therefore it may be the case that $\mathcal{D} \models \phi(\sigma_{\overline{Cl(\alpha_n,\sigma_n)}})$.

For example, consider the new trace $\sigma_{cc2} = do([teleHack(H_2,C), recover(C), drive(C,I_1,I_2), keyHack(H_1,C), teleHack(H_2,C)], S_0)$. Although intuition suggests that $\{(drive(C,I_1,I_2),S_2'),(keyHack(H_1,C),S_3')\}$ (where $S_2' = do([teleHack(H_2,C), recover(C)], S_0)$ and $S_3' = do(drive(C,I_1,I_2),S_2')$) is an actual achievement cause for the effect ϕ_{cc}, the effect ϕ_{cc} is not counterfactually dependent on this actual contributing cause, since removing the final action of this cause produces the trace $[teleHack(H_2,C), recover(C), drive(C,I_1,I_2), teleHack(H_2,C)]$, whose execution starting in S_0 brings about ϕ_{cc} anyway.

4 Actual Achievement Causes

In this section, we identify a simple property that is both necessary and sufficient for actual achievement causes. Based on this, we then give our new definition of actual cause. We start by formulating a more inclusive sufficient condition (than in the previous section). In this, we want to ensure that the cause $\mathcal{K}$ is not preempted by another actual contributing cause. The definition must also take into consideration any possible contributing cause that is preempted by $\mathcal{K}$. Finally, we must ensure that the effect of $\mathcal{K}$ is enduring.

Definition 11 (Weak Sufficiency). *Given causal setting $\mathcal{C} = \langle\mathcal{D},\sigma,\phi(s)\rangle$ and an ordered set of (action, situation) pairs $\mathcal{K} = \{(\alpha_1,\sigma_1),\cdots,(\alpha_n,\sigma_n)\}$ taken from σ. We say that $\mathcal{K}$ is weakly sufficient for ϕ if and only if $\mathcal{K}$ is an actual contributing cause of $\mathcal{C}$ with witness $\{\sigma_1,\cdots,\sigma_n\}$, and*

$$\mathcal{D} \models \neg\phi(\sigma_n).$$

Notice the difference from Def. 10, where the effect $\neg\phi$ was verified at the end of $\sigma_{\overline{Cl(\alpha_n,\sigma_n)}}$; in contrast, in Def. 11, $\neg\phi$ is verified

[6] Recall that unlike $<$ and $\leq$, the precedence operators $\sqsubset$ and $\sqsubseteq$ do not require executability. Also, here "Cl" means "closure".

in situation σ_n. By doing this, we omit the effect of the final action α_n in $\mathcal{K}$, and the effects of all subsequent actions that occur on the trace after the situation σ_n where α_n is performed. This is important since if there is a competing enduring cause that was preempted by $\mathcal{K}$, it can manifest only after the situation σ_n. By omitting all such situations (and actions performed in those situations) from our consideration in Def. 11 we are essentially ignoring the effects of all potential competing causes that are preempted by $\mathcal{K}$.

Note that if ϕ counterfactually depends on $\mathcal{K}$, then $\mathcal{K}$ is weakly sufficient for ϕ, but not vice-versa. In this sense, weak sufficiency is a weaker condition than counterfactual dependence w.r.t actual contributing causes. It is important to note that weak sufficiency by itself *does not* guarantee causation, since a subsequent action from σ – one that occurs after situation σ_n – may render the effect false. However, in the absence of such subsequent actions (i.e. when $\mathcal{K}$ is also an enduring producer of ϕ given causal setting $\mathcal{C}$), one can in fact guarantee that if $\mathcal{K}$ is weakly sufficient for ϕ, then $\mathcal{K}$ is also sufficient for ϕ, since $\mathcal{K}$ is enduring, i.e. ϕ remains true from situation $do(\alpha_n, \sigma_n)$ up to the end of the trace σ. Thus weak sufficiency, along with endurance of the achieved effect, is a sufficient condition for causation.

In our original example (with trace σ_{cc}), the only enduring producer $\{(drive(C, I_1, I_2), S_3), \quad (keyHack(H_1, C), S_4)\}$ is weakly sufficient for ϕ_{cc}. Moreover, in our modified example with trace σ_{cc2}, the enduring producer $\mathcal{K}_{cc2} = \{(drive(C, I_1, I_2), S_2'), (keyHack(H_1, C), S_3')\}$ is weakly sufficient for ϕ_{cc}; recall that ϕ_{cc} is not counterfactually dependent on $\mathcal{K}_{cc2}$.

It can be shown that enduring producers are weakly sufficient.

Proposition 4 (Enduring Producers are Weakly Sufficient). *Let $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s)\rangle$ be a causal setting and $\mathcal{K}$ be an enduring producer of $\mathcal{C}$. Then $\mathcal{K}$ is weakly sufficient for ϕ given setting $\mathcal{C}$.*

Put otherwise, if $\mathcal{K} = \{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ is an enduring producer of the causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s)\rangle$, then $\mathcal{D} \models \neg\phi(\sigma_n)$, i.e. the theory $\mathcal{D}$ entails that the effect ϕ cannot be observed in the situation obtained by executing all the actions of σ except for those that occur after situation σ_n, starting from the initial situation S_0. Thus enduring production is indeed a property that is **both necessary and sufficient** for actual achievement causes. It is a necessary property in the sense of Propositions 1, 2, and 3. That is, each part of an enduring producer must contribute to the achievement of the effect. An enduring producer can not be preempted, and the effect brought about by an enduring producer must persist. It is a sufficient property for actual achievement causes in the sense of Proposition 4. That is, had the final action in the enduring producer not occurred, the effect ϕ would not have been observed, under the contingency that all subsequent competing (but preempted) causes are ignored.

We are now ready to give our new definition. We define actual achievement causes simply as enduring producers of the effect.

Definition 12 (Actual Achievement Cause). *Given a causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s)\rangle$, a non-empty ordered set of action-situation pairs $\{(\alpha_1, \sigma_1), \cdots, (\alpha_n, \sigma_n)\}$ is an actual cause of $\phi(s)$ if and only if it is an enduring producer of $\phi(s)$ given $\mathcal{C}$.*

Implementation. We developed a preliminary implementation of our definition of actual cause based on Reiter's regression [35]. We used the simpler definition in [1] for this (thanks to Corollary 1, see below). The current version does not handle quantified effects, and it can be improved. We tested queries from the example mentioned in the paper, as well as causal queries from a simple blocks world domain with about 20 blocks. Our program computed actual causes in less than 0.02 seconds, i.e., that our implementation is quite efficient.

5 The Batusov-Soutchanski (2018) Approach

According to Batusov and Soutchanski [1], if some action α of the action sequence in σ triggers the formula $\phi(s)$ to change its truth value from false to true relative to $\mathcal{D}$, and if there are no actions in σ after α that change the value of $\phi(s)$ back to false, then α is an actual cause of achieving $\phi(s)$ in σ. They showed that when used together with the single-step regression operator ρ, in addition to the single action that brings about the effect of interest, one can also capture the chain of actions that build up to it. The following inductive definition formalizes this intuition. Let $\Pi_{apa}(\alpha, \sigma)$ be the right-hand side of the precondition axiom for action α in situation σ.

Definition 13 (Actual Achievement Cause (Batusov-Soutchanski 2018)). *A causal setting $\mathcal{C} = \langle \sigma, \phi(s)\rangle$ satisfies the achievement condition of ϕ via the situation term $do(\alpha^*, \sigma^*) \sqsubseteq \sigma$ if and only if there is an action α' and situation σ' such that*

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

and either $\alpha^* = \alpha'$ and $\sigma^* = \sigma'$, or the causal setting $\langle \sigma', \rho[\phi(s), \alpha'] \wedge \Pi_{apa}(\alpha', \sigma')\rangle$ satisfies the achievement condition via the situation term $do(\alpha^*, \sigma^*)$. Whenever a causal setting $\mathcal{C}$ satisfies the achievement condition via situation $do(\alpha^*, \sigma^*)$, the action α^* executed in situation σ^* is said to be an achievement cause in $\mathcal{C}$.

Batusov and Soutchanski [1] show that the achievement causes of $\mathcal{C}$ form a finite sequence of situation-action pairs, which they call the *achievement causal chain of $\mathcal{C}$*.

Formal Relationship. We can now show that our definition is indeed equivalent to the one proposed in [1]. In particular, counterfactual dependence entails their achievement causes.

Theorem 1. *Given setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s)\rangle$, if an effect ϕ is counterfactually dependent on some action α^* with witness σ^*, then α^* executed in σ^* is an achievement cause of ϕ in $\mathcal{C}$ according to [1].*

Proof Sketch. *Fix $\mathcal{C} = \langle \sigma, \phi\rangle, \alpha^*$, and σ^*, and assume that the antecedent is true, i.e. that ϕ counterfactually depends on α^* executed in σ^*. From this and Def. 9, we have: $\mathcal{D} \models \neg\phi(\sigma_{\overline{Cl(\alpha^*, \sigma^*)}})$. By Def. 1, $\mathcal{D} \models \phi(\sigma)$. This gives us 2 cases: (Case-1). There is some action α' executed in some situation σ' s.t. $(\alpha', \sigma') \in \langle Cl(\alpha^*, \sigma^*, \sigma)\rangle$ and s.t. α' achieves ϕ in σ. Then by Def. 13, α' executed in σ' is a cause of ϕ. Also, by this and Def. 13, any action executed in some situation between S_0 and σ' that directly, and by induction indirectly, brings about the preconditions of α' must be a cause of ϕ. Since (α', σ') is in $\langle Cl(\alpha^*, \sigma^*, \sigma)\rangle$, the preconditions of α' depends on α^*, and thus (α^*, σ^*) is a cause. (Case-2). There is some action α' executed in situation σ' that achieves ϕ in σ, but $(\alpha', \sigma') \notin \langle Cl(\alpha^*, \sigma^*, \sigma)\rangle$. By this and by the antecedent, α' executed in σ' does not achieve ϕ in $\sigma_{\overline{Cl(\alpha^*, \sigma^*)}}$. This can only happen if the removal of the actions in $\langle Cl(\alpha^*, \sigma^*, \sigma)\rangle$ removed some condition ψ that is required for the achievement of ϕ when α' is executed in σ'. By this, Def. 13, and by induction, there is an action, say α'' executed in σ'', s.t. $(\alpha'', \sigma'') \in \langle Cl(\alpha^*, \sigma^*, \sigma)\rangle$ and $S_0 \leq \sigma'' < \sigma'$, whose execution brings about ψ, and (α'', σ'') is a cause of the effect ϕ. Moreover,*

by induction, it can be shown that since the preconditions of α'' directly or indirectly depends on α^* (as $(\alpha'', \sigma'') \in \langle Cl(\alpha^*, \sigma^*, \sigma)\rangle$), (α^*, σ^*) is also a cause of ϕ. □

Their causal chains are enduring producers and vice-versa.

Theorem 2. *Given setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s)\rangle$, $\mathcal{K}$ is a causal chain relative to $\mathcal{C}$ according to [1] if and only if $\mathcal{K}$ is an enduring producer of ϕ given $\mathcal{C}$.*

Proof Sketch. $(=>)$ *The proof involves showing that the primary cause w.r.t a setting $\mathcal{C}$ is unique; this follows from Def. 13, the fact that the trace is linear and thus there can be only one action α^* and situation σ^* on the trace s.t. $\mathcal{D} \models \neg\phi(\sigma^*) \wedge \phi(do(\alpha^*, \sigma^*)) \wedge \forall s'(do(\alpha^*, \sigma^*) \leq s' \leq \sigma) \rightarrow \phi(s')$, and that the output of ρ is unique. The rest of the proof involves constructing an enduring producer that matches the causal chain, starting from the final situation of the chain. This can be done by showing by induction on the length of the causal chain that all (action,situation) pairs in the causal chain are on the enduring producer, and showing that the sequence is maximal (i.e. it can't be extended in the past).* $(<=)$ *This case is similar, but uses Prop. 2 to show an enduring producer is unique.* □

Consequently, our causes are equivalent to their causal chains.

Corollary 1. *Given causal setting $\mathcal{C} = \langle \mathcal{D}, \sigma, \phi(s)\rangle$, $\mathcal{K}$ is a causal chain relative to $\mathcal{C}$ according to [1] if and only if $\mathcal{K}$ is an actual achievement cause of $\mathcal{C}$.*

6 Discussion

Following Hume's definition and motivated by the Lewis's [26] paper, there has been much work investigating the relationship between causality and counterfactual reasoning [31]. Researchers have argued that reasoning about counterfactual worlds plays an indispensable role when determining causation [32]. Experimental results from psychology show that varying relevant counterfactual worlds while keeping the actual world events fixed strongly affect participants' causal judgments [8]. In contrast, keeping the counterfactual worlds constant and varying how the actual outcome was brought about much less influence their causal judgments. This demonstrates that human causal judgments are indeed inextricably linked to counterfactuals. Researchers have emphasized the close interrelation between causality and counterfactuals while studying causal responsibility [24] and causation in legal and moral reasoning [23].

The HP Approach. Of particular note is Halpern and Pearl's inspiring work on actual causality. In [33], Pearl proposed a definition of actual cause based on the notion of causal beams. An improved version appeared in the first edition of Pearl's book [34]. However, it handled path switching examples incorrectly [37, 1]. As a remedy, HP introduced their original definition in [17]. Counter examples by Hopkins and Pearl [21] motivated the updated definition, which was introduced in [18]. However, their updated definition did not agree with intuition as was shown using counter examples by Hopkins [20], Weslake [37], and others. To deal with this, the modified definition was introduced in [20, 15]. Unfortunately, this latest definition also has issues (see below). In the words of Halpern ([16], p.27) "The jury is still out on what the 'right' definition of causality is".

HP's approach is based on the framework of Structural Equations Models (SEM). An acyclic SEM model consists of an ordered set of assignments processed top down, where each endogenous variable on the left takes a value computed using a function. The function is such that its arguments are values of other variables computed from the preceding equations. The definition of actual cause is given using interventions that allow the overriding of the values of some variables to model counterfactuals. In [1], it is shown that SEM and interventions can be formulated in terms of a basic action theory in the SC.

Due to lack of space we cannot go over the details of all three HP definitions, but the technical details are well presented in [16, 37] and elsewhere. However, we would like to discuss the modified definition [16], since it attempts to address the problems in the previous definitions. Notice we do not argue against SEM approach in general.

Let $\mathcal{U}$ and $\mathcal{V}$ be the sets of exogenous and endogenous variables, $(M, \bar{V}_\mathcal{U})$ be a causal setting, X be an endogenous variable, and V_X be the value of X, see [16] for details. The conjunction of primitive events $\bar{X} = \bar{V}_X$, short for $X_1 = V_{X_1} \wedge \cdots \wedge X_k = V_{X_k}$, is an actual cause in $(M, \bar{V}_\mathcal{U})$ of a HP query ϕ if all the following conditions hold: **1.** $(M, \bar{V}_\mathcal{U}) \models (\bar{X} = \bar{V}_X)$ and $(M, \bar{V}_\mathcal{U}) \models \phi$. **2.** There exists a set $\bar{W}$ (disjoint from $\bar{X}$) of variables in $\mathcal{V}$ with $(M, \bar{V}_\mathcal{U}) \models (\bar{W} = \bar{V}_W)$ and a setting $\bar{V}'_X$ of variables $\bar{X}$ such that $(M, \bar{V}_\mathcal{U}) \models [\bar{X} \leftarrow \bar{V}'_X, \bar{W} \leftarrow \bar{V}_W]\neg\phi$. **3.** No proper subconjunction of $(\bar{X} = \bar{V}_X)$ satisfies 1, 2. 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. Note that in Item 2, according to $(M, \bar{V}_\mathcal{U}) \models (\bar{W} = \bar{V}_W)$, interventions that set variables in $\bar{X}$ to counterfactual values $\bar{V}'_X$ must set all variables in $\bar{W}$ to their actual values $\bar{V}_W$ in the actual context. This means, according to the modified definition, that if the set $\bar{W} \neq \emptyset$, then the values assigned to exogenous variables and/or the counterfactual values $\bar{V}'_X$ cannot propagate downstream in the set of equations to influence the values of $\bar{W}$ and the values of those variables which directly or indirectly depend on $\bar{W}$. Contrary to the constraints embodied in the SEM, only selective propagation is allowed since values of variables $\bar{W}$ are fixed by the actual context. This is counter-intuitive since each equation reflects a mechanism in a model. Hitchcock [19] labelled such selective propagations *explicitly nonforetracking counterfactuals*. Hall [13] pointed out that the analysis of actual causation using non-actual worlds where the causal relations do not hold is non-intuitive. Halpern acknowledges this, by calling such interventions "miraculous" ([16], p.32), and introduces the notion of normality in an attempt to minimize them.

To illustrate the problem with the modified definition, we use the well known "bottle" example [12]. Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first shattering the bottle. Billy's rock would have shattered it had it not for Suzy's. The story can be modeled using the following set of structural equations that uses 5 endogenous variables: $ST := 1, BT := 1, SH := ST, BH := BT \wedge \neg SH, BS := SH \vee BH$, where ST, SH, BT, BH, and BS stand for Suzy throws, Suzy hits, Billy throws, Billy hits, and bottle shatters. Note that, according to the modified definition, ST is the cause of BS since taking a witness $BH := 0$ (Billy did not hit) yields: $(M, \bar{V}_\mathcal{U}) \models [ST \leftarrow 0, BH \leftarrow 0]\neg BS$. However, this reasoning is counter-intuitive. This counterfactual is physically impossible and violates the model. Namely, if $ST := 0$ (Suzy did not throw) and $BT := 1$ (Billy did throw), then BH must be 1 according to the given equations, not 0. Notice also that BS equals to $BT \vee ST$; so the conclusions of this definition depend on the syntax. The underlying issue of this counter-intuitive argument is in selective propagation of values from interventions to the endogenous variables. We sidestep this conceptual problem with the modified definition, since we propose a new definition of actual achievement causes that avoids selective propagation altogether. To see why the aforementioned issues are not a problem in our defini-

tion, it is sufficient to notice that our basic action theory $\mathcal{D}$ is fixed and therefore valuations such as $\neg ST \wedge BT \wedge \neg BH$ cannot occur.

In the following, we show how the bottle example can be formalized in our framework. Here, we introduce two additional actions, each representing a hit on the bottle. This abstracts away from the time of throwing a rock and the duration after which it hits the bottle.[7] The actions in our formalization of the bottle example are thus $throws(p)$ and $hits(p)$, where p is either $Suzy$ or $Billy$. The preconditions for these actions are as follows:

$$\forall p, s.\ (Poss(throws(p), s) \leftrightarrow true),$$
$$\forall p, s.\ (Poss(hits(p), s) \leftrightarrow thrown(p, s)).$$

There are two fluents, $thrown(p, s)$, representing that person p has thrown the rock in situation s, and $broken(s)$, that the bottle is broken in s. Their successor-state axioms are as follows:

$$\forall p, a, s.\ thrown(p, do(a, s)) \leftrightarrow (\ a{=}throw(p) \vee thrown(p, s)\ ),$$
$$\forall a, s.\ broken(do(a, s)) \leftrightarrow (\ \exists p(a{=}hits(p)) \vee broken(s)\ ).$$

Initially, the bottle is intact and no rocks have been thrown, i.e. $\forall p.(\neg thrown(p, S_0))$ and $\neg broken(S_0)$. Let us now consider three causal settings $\langle do(\vec{\alpha}, S_0), broken(s)\rangle$, where $\vec{\alpha}$ can be as follows:

$i.\ [throws(Suzy), throws(Billy), hits(Suzy), hits(Billy)]$,
$ii.\ [throws(Billy), throws(Suzy), hits(Suzy), hits(Billy)]$,
$iii.\ [throws(Billy), throws(Suzy), hits(Billy), hits(Suzy)]$.

By Definition 12, the causes relative to these settings are $\{(throws(Suzy), S_0), (hits(Suzy), S_2^i)\}$, $\{(throws(Suzy), S_1^{ii}), (hits(Suzy), S_2^{ii})\}$, and $\{(throws(Billy), S_0), (hits(Billy), S_2^{iii})\}$, respectively. Here $S_2^i = do([throws(Suzy), throws(Billy)], S_0)$, $S_1^{ii} = do(throws(Billy), S_0)$, $S_2^{ii} = do(throws(Suzy, S_1^{ii}))$, etc. Note that, as expected, Billy's throw and subsequent hit is the cause in the last causal setting where his throw hits the bottle first.

The INUS Condition. Interestingly, our new definition of actual cause can be used to illustrate Mackie's [27] account of Hume's regularity definition. Elaborating on Hume, Mackie proposed the so called INUS condition, which postulates that A is a cause of B if A is an Insufficient but Necessary part of a condition that is itself Unnecessary but Sufficient for B. Put otherwise, A is a cause of B if there exist X and Y such that $(A \wedge X) \vee Y$ is both necessary and sufficient for B, but neither A nor X by itself is sufficient to entail B. For simplicity, we present an argument involving causes that consist of a single action-situation pair only, or for that matter, primary causes. The argument for causes with multiple action-situation pairs is similar. Let us illustrate the INUS condition relative to our definition starting from right (*S*) to left (*I*). First note that an enduring producer is *S*ufficient for causation. This is because the existence of the enduring producer on the trace guarantees causation. But it is *U*nnecessary for causation, since other subsequent (preempted) actions could have brought about the effect had the enduring producer not occurred.

Next, we will show that the actual cause itself is in some sense an insufficient but necessary part of the enduring producer. To see this, note that while a (primary) cause is a pair (α^*, σ^*), we use the situation σ^* only to uniquely identify which instance of action α^* is a cause, since the trace may include multiple occurrences of α^*. In other words, if all actions had distinct names, we could have ignored σ^*. Thus, a (primary) cause is essentially an action α^* in our framework, one that represents the *N*ecessary part of our unnecessary but sufficient condition (i.e. of the enduring producer). The action α^* executed in σ^* is necessary since the effect ϕ was false in σ^*, and ϕ is achieved only after the execution of α^* in σ^*. Finally, α^* by itself is *I*nsufficient, since the execution of α^* in some arbitrary situation does not guarantee the achievement of the effect ϕ. Rather, α^* only achieves ϕ under some appropriate conditions, which in our case, are captured by the situation σ^*.

Put another way, $(A \wedge X)$ can be understood as the enduring producer, Y includes any subsequent (preempted) actions that occur on the trace after the enduring producer, A is the action part of the enduring producer, and X includes the conditions under which the execution of the actions in the enduring producer achieve ϕ, namely those that are captured by the situations on the trace where these actions are executed. Note that, we illustrate Mackie's approach not w.r.t general causality, but w.r.t actual causality. While the original INUS definition was criticized because it does not deal with preemption, we avoid this problem by taking into account the order in which actions occurred in the trace. For us, only the earliest actual enduring contributing cause on the trace can be an actual cause. Therefore, the actual cause could not have been preempted by another contributing cause. Notice our illustration of Mackie's proposal transcends critical comments in Section 10.1.4 of J. Pearl's book [34].

Other Approaches. Our notions of contributors and producers resemble that of discussed in [3] and [12], but the connections end there. In particular, our action framework, i.e. the SC, is much more expressive than theirs. With this expressiveness, we encounter subtleties that must be dealt with. For example, we now need to deal with objects, preconditions of actions, and the non-persistence of fluents, among other things. As effects are simply events in their account, once they occur, they cannot "unoccur" again. On the other hand, effects in the situation calculus are uniform formulae consisting of fluents, and fluents can change their value depending on the situation. Thus, effects in our framework can be undone by other events, and only the first "enduring" actual contributing cause can be considered as an actual cause. Although by incorporating a temporal order on events into the equation, [3] made a positive step towards an acceptable solution (roughly by ensuring that the event that achieved the effect is the earliest occurring one), their model of this is somewhat ad-hoc and relies on some unusual notions, e.g. that of time of "occurrence" of an absent event, etc. On the other hand, we use a formal language of action and change to define actual causation, where timing (i.e., ordering) of actions or events is implicitly obtained from the underlying situations; e.g. if $S_0 \leq do(\alpha, \sigma_1) < do(\beta, \sigma_2) \leq \sigma$, then event α must have occurred before event β.

Previously, Hopkins and Pearl [22] attempted to capture counterfactuals within the SC. However, they did not take preconditions of actions into account. They did not define actual cause in their paper.

Recently, Bochman [5, 6] gave a definition of actual cause in the causal calculus [4], a non-monotonic formalism introduced by McCain and Turner [28] for reasoning about actions. His definition is based on a causal version of the INUS condition, namely the NESS condition [38]. While our definition of actual achievement cause can informally illustrate Mackie's approach, our definition is motivated and derived from a very different perspective. Also, as we explained, our definition cannot be reduced to any arbitrary INUS condition. Rather, our actual cause is an appropriately selected INUS condition. Namely, this condition to be right must depend on the order and occurrence of actions in the trace. On the other hand, the definition given in [5] is based on the regularity approach, and it is, to quote from [5], "a direct formalization of the NESS test".

In [5] and [6], causation is defined between propositions. In con-

[7] As can be seen above, Halpern and Pearl also abstracts away from the temporal aspects by introducing variables BH and SH.

trast, our ontology for causes and effects is different. Namely, our causes are actions (or events) and effects are (uniform) formulae in the situation calculus. Moreover, our framework builds on the more expressive situation calculus with a standard first-order logic semantics in contrast to propositional non-monotonic causal calculus in [5, 6] that has a non-standard semantics. Therefore, unlike [5] and [6], we can deal with effects formulated as quantified formulas. Furthermore, we can compute causal chains in our framework using one-step regression. It is not obvious how this can be done in the non-monotonic causal calculus.

In [6], Bochman argued that actual causes can be defined without appeal to counterfactuals, which are shown, using examples, to have their limitations within the causal theories. As stated above, we consider counterfactuals within the situation calculus, which has a different ontology and expressiveness than causal theories. Whether the limitations of counterfactuals mentioned in [6] also manifest in other frameworks, including our approach, is an open question.

The event calculus [30] is another well-known formalism for reasoning about events and change. But since it does not include situations, and therefore the regression operator cannot be defined, it is not clear how one can use the event calculus to replicate our approach.

7 Conclusion and Future Work

In this paper, we studied the intuitive properties that are necessary for actual causes and conditions that are sufficient for the achievement of an observed (possibly quantified) effect. We identify a property that is both necessary and sufficient for actual achievement causes. This lead to a new definition of actual achievement causes. We prove that our new definition is equivalent to the one proposed by Batusov and Soutchanski [1]. This shows that their foundational definition of actual achievement causes can be understood in counterfactual terms. Also, our definition can illustrate Mackie's interpretation of the regularity account. Thus, we contribute to the long standing debate between the regularity and the counterfactual camps by giving a definition that was derived using counterfactual analysis and that can illustrate Mackie's definition. In some sense, our paper contributes towards the task of bridging the gap between these two camps. As in [1], we focused on linear traces only. In the future, we would like to study cases where the order of actions is given only partially.

Acknowledgements. We thank the anonymous reviewers for helpful comments. This work was supported in part by the NSERC Canada and by the Faculty of Science at Ryerson University.

REFERENCES

[1] V. Batusov and M. Soutchanski, 'Situation calculus semantics for actual causality', in *Proc. AAAI*, pp. 1744–1752, (2018).
[2] M. Baumgartner, 'A regularity theoretic approach to actual causation', *Erkenntnis*, **78**(1), 85–109, (2013).
[3] S. Beckers and J. Vennekens, 'A principled approach to defining actual causation', *Synthese*, **195**(2), 835–862, (2018).
[4] A. Bochman, 'A logic for causal reasoning', in *Proc. IJCAI*, pp. 141–146, (2003).
[5] A. Bochman, 'Actual causality in a logical setting', in *Proc. IJCAI*, pp. 1730–1736, (2018).
[6] A. Bochman, 'On laws and counterfactuals in causal reasoning', in *Proc. KR*, pp. 494–503, (2018).
[7] T. Eiter and T. Lukasiewicz, 'Complexity results for structure-based causality', *Artificial Intelligence*, **142**(1), 53–89, (2002).
[8] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, 'From counterfactual simulation to causal judgment', in *Proc. Cognitive Science Society*, pp. 523–528, (2014).
[9] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, 'How, whether, why: Causal judgments as counterfactual contrasts', in *Proc. Cognitive Science Society*, pp. 782–787, (2015).
[10] C. Glymour, D. Danks, B. Glymour, F. Eberhardt, J. Ramsey, R. Scheines, P. Spirtes, C. Man Teng, and J. Zhang, 'Actual causation: A stone soup essay', *Synthese*, **175**(2), 169–192, (2010).
[11] A. Greenberg. Hackers remotely kill a Jeep on the highway – with me in it, 2015. https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway, Retrieved on August 25, 2019.
[12] N. Hall, 'Two concepts of causation', in *Causation and Counterfactuals*, eds., J. Collins, N. Hall, and L. Paul, 225–276, MIT Press, (2004).
[13] N. Hall, 'Structural equations and causation', *Philosophical Studies*, **132**, (01 2007).
[14] J. Y. Halpern, 'Axiomatizing causal reasoning', *J. of Artificial Intelligence Research*, **12**, 317–337, (2000).
[15] J. Y. Halpern, 'A modification of the Halpern-Pearl definition of causality', in *Proc. IJCAI*, pp. 3022–3033, (2015).
[16] J. Y. Halpern, *Actual Causality*, The MIT Press, 2016.
[17] J. Y. Halpern and J. Pearl, 'Causes and explanations: A structural-model approach: Part 1: Causes', in *Proc. UAI*, pp. 194–202, (2001).
[18] J. Y. Halpern and J. Pearl, 'Causes and explanations: A structural-model approach. Part I: Causes', *The British J. for the Philosophy of Science*, **56**(4), 843–887, (2005).
[19] C. Hitchcock, 'The intransitivity of causation revealed in equations and graphs', *J. of Philosophy*, **98**(6), 273–299, (2001).
[20] M. Hopkins, *The Actual Cause: From Intuition to Automation*, Ph.D. dissertation, Univ. of California at Los Angeles, 2005.
[21] M. Hopkins and J. Pearl, 'Clarifying the usage of structural models for commonsense causal reasoning', in *In Proc. Commonsense*, (2003).
[22] M. Hopkins and J. Pearl, 'Causality and counterfactuals in the situation calculus', *J. of Logic and Computation*, **17**(5), 939–953, (2007).
[23] D. A. Lagnado and T. Gerstenberg, 'Causation in legal and moral reasoning', *Oxford Handbook of Causal Reasoning*, 565–602, (2017).
[24] D. A. Lagnado, T. Gerstenberg, and R. Zultan, 'Causal responsibility and counterfactuals', *Cognitive Science*, **37**(6), 1036–1073, (2013).
[25] F. Leitner-Fischer and S. Leue, 'Causality checking for complex system models', in *Proc. 14th VMCAI*, LNCS, vol.7737, pp. 248–267, (2013).
[26] D. Lewis, 'Causation', *The J. of Philosophy*, **70**(17), 556–567, (1974).
[27] J. L. Mackie, 'Causes and conditions', *American Philosophical Quarterly*, v.**2**(4), 245–264, (1965).
[28] N. McCain and H. Turner, 'Causal theories of action and change', in *Proc. AAAI*, pp. 460–465, (1997).
[29] J. McCarthy and P. J. Hayes, 'Some philosophical problems from the standpoint of artificial intelligence', *Mach. Intell.*, v.**4**, 463–502, (1969).
[30] E. T. Mueller, *Commonsense Reasoning: An Event Calculus Based Approach*, Morgan Kaufmann, 2014.
[31] P. Menzies and H. Beebee, 'Counterfactual Theories of Causation', Stanford Encyclopedia of Philosophy, `https://plato.stanford.edu/entries/causation-counterfactual/`, 2019, Retrieved on February 14, 2020.
[32] C. L. Ortiz, 'Explanatory update theory: Applications of counterfactual reasoning to causation', *Artif. Intell.*, v. **108**(1-2), 125–178, (1999).
[33] J. Pearl, 'On the definition of actual cause', Technical report, R-259, University of California L.A., (1998).
[34] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000.
[35] R. Reiter, *Knowledge in Action. Logical Foundations for Specifying and Implementing Dynamical Systems*, MIT Press, 2001.
[36] H. A. Simon, 'Causal ordering and identifiability', *Models of Discovery. Boston Studies in the Philosophy of Science*, **54**, (1977).
[37] B. Weslake, 'A partial theory of actual causation', *British J. for the Philosophy of Science*, (2015).
[38] R. W. Wright, 'Causation in tort law', *California Law Review*, **73**(6), 1735–1828, (1985).