What Is a Counterfactual Cause in Action Theories?

Extended Abstract

Daxin Liu
Nanjing University
Nanjing, China
daxin.liu@nju.edu.cn

Vaishak Belle
The University of Edinburgh
Edinburgh, United Kingdom
vbelle@ed.ac.uk

ABSTRACT

Since the proposal by Halpern and Pearl, reasoning about actual causality has gained increasing attention in artificial intelligence, ranging from domains such as model-checking and verification to reasoning about actions and knowledge. More recently, Batusov and Soutchanski proposed a notion of actual achievement cause in the situation calculus, amongst others, they can determine the cause of quantified effects in a given action history. While intuitively appealing, this notion of cause is not defined in a counterfactual perspective. In this paper, we propose a notion of cause based on counterfactual analysis. In the context of action history, we show that our notion of cause generalizes naturally to a notion of achievement cause.

KEYWORDS

Reasoning about actions, Counterfactual causality, Causality in actions

ACM Reference Format:
Daxin Liu and Vaishak Belle. 2025. What Is a Counterfactual Cause in Action Theories?: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Causality [13] plays a central role in artificial intelligence by determining how an agent understands its observations. The topic has evolved across different sub-communities: knowledge representation typically focuses on actions and effects [14], while philosophers and machine learning researchers explore type/general causality (e.g., does smoking cause cancer) and actual causality [4] (e.g., whether the harm to the victim was carried out by the perpetrator or an accidental fire in the victim's home). Actual causation, as argued by Halpern and Hitchcock [5], remains contentious due to competing formalisms. It builds on David Hume's "but-for" causality [7], which has been formalized through structural equation models [12] and refined in the Halpern-Pearl (HP) account [3].

While structural equation models offer an attractive framework, their simplicity makes domain modeling challenging [4, 6]. This has led to alternative approaches in the situation calculus, most recently by Batusov and Soutchanski [1]. Their key idea is to provide a definition of "achievement cause" which, given a history of actions and an outcome, tries to use regression to identify the subsequence that leads to this event being true. Although this is a very worthy start, what we ask in this paper is whether it is possible to characterize actual causality simply in terms of the minimal conditions for an event being true. We argue that a simple definition that almost completely lifts the "but-for" definition of causality, but in a rich action setting, is possible. Essentially, our definition follows that of Lewis's intuition, but also that of the HP modified definition using structural equation models [3].

Though our approach offers uniform standard modeling through action theories [14], it faces limitations with disjunctive goals and interleaving actions. Essentially, although the identification of cause in our proposal may seem counterintuitive in such examples, this is simply an artifact of the domain itself and requires us to think carefully about the notion of actual cause with such goals. We do not think this necessarily is definitive proof that one formalism is better than the other. Rather, we believe the situation calculus [14] and its simple ontology do provide a very natural way to think about actions, effects, preconditions, and the role they could play in actual causation, with promising applications in robot programming through GOLOG [10]. We provide more details on these aspects, including a detailed discussion of the HP account in an extended report [11].

2 A MODAL LOGIC OF ACTION AND CHANGE

We use the logic $\mathcal{ES}$ [9] to model actions and change, a modal variant of the situation calculus. The logic features a fixed countable domain called *standard names* which amounts to having an infinite domain closure axiom together with the unique name assumption.

Syntax. The logic has two sorts: *object* and *action*. The vocabulary includes the usual gradients of first-order logic, together with two modalities $[\cdot], \Box$. $[t]\phi$ and $\Box\phi$ are read as ϕ holds after action t and after any action sequence, respectively. For action sequence $z = a_1 \cdots a_k$, we write $[z]\alpha$ to mean $[a_1] \ldots [a_k]\alpha$. The logic includes a special fluent $Poss(a)$ to express action a is executable.

Semantics. The semantics is given in terms of possible worlds and a world w determines what holds initially and after any action sequences. Namely, a world w maps every *primitive formula* $F(n_1, \ldots, n_k)$ and *action sequence* z to $\{0, 1\}$. Let $\mathcal{Z}$ be the set of action sequences (including the empty sequence $\langle\rangle$).

Definition 2.1 (Truth of Formulas). Given a world $w \in \mathcal{W}$ (the set of all worlds) and a sentence ψ, we define $w \models \psi$ as $w, \langle\rangle \models \psi$, where for any $z \in \mathcal{Z}$: (negation, connectives, and quantifiers are handled in the usual sense)

- $w, z \models F(n_1, \ldots, n_k)$ iff $w[F(n_1, \ldots, n_k), z] = 1$;
- $w, z \models [t]\psi$ iff $w, z \cdot t \models \psi$;

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

- $w, z \models \Box\psi$ iff $w, z \cdot z' \models \psi$ for all $z' \in Z$;

Satisfiability, validity, and logical entailment are defined as usual.

Basic Action Theory. $\mathcal{ES}$ use a variant of the *basic action theory* (BAT Σ) to express the dynamic of a domain, which contains the initial state axiom Σ_0, the action preconditions axiom Σ_{ap} and the successor state axioms Σ_{post}. E.g., the following $\Sigma_{ap}, \Sigma_{post}$ specify a block domain [8]: [1]

$$\begin{aligned} &\Box Poss(a) \equiv (a = pickup(x) \wedge \neg Holding(x)) \\ &\qquad \vee (a = drop(x) \wedge Holding(x)) \\ &\Box[a]Holding(x) \equiv a = pickup(x) \\ &\qquad \vee a \neq drop(x) \wedge Holding(x) \\ &\Box[a]Broken(x) \equiv a = drop(x) \vee Broken(x) \end{aligned} \tag{1}$$

That is picking up is always possible and dropping is only possible when it is already holding the object. Moreover, *Holding* might be affected by the action *pickup* and *drop* in the literal way. *drop* an object might cause it to be broken. Suppose Σ_0 is as $\{\neg Holding(x), \neg Broken(C), \neg Broken(D)\}$, then

$$\Sigma \models [pickup(C)]\,(Holding(C) \wedge [drop(C)]Broken(C))$$

To ensure that the action sequence is executable, we define *exec* as $exec(\langle\rangle) = \textsc{True}$, and $exec(a \cdot z) = Poss(a) \wedge [a]exec(z)$.

3 COUNTERFACTUAL ACHIEVEMENT CAUSE

Minimal Cause. We start with the notion of minimal cause which is based on the notion of counterfactual.

Definition 3.1 (Minimal cause). Given a BAT Σ and a static sentence ϕ representing the goal, we said an action sequence z is the *minimal cause* of ϕ wrt Σ if (1) $\Sigma \models \neg\phi$; (2) $\Sigma \models [z]\phi$; (3) z is minimal.

Clearly, this definition requires a notion of distance between action sequences and $\langle\rangle$. One could easily define causality in terms of *length*, *affected fluent*, and so on, of action sequences [2]. We note that this notion of minimal causes is counterfactual: initially, goal ϕ does not hold factually. When considering all alternative situations, had the action sequences, i.e. the cause, not happened, the goal would not have been achieved via another action sequence that is *smaller* than the cause.

In certain scenarios, one might know the action history, i.e. the so-called *narratives*, and wish to find the exact actions that cause the goal. This is exactly what Batusov and Soutchanski [1] do. Here, we propose an alternative in terms of counterfactual:

Minimal Cause in Narratives. By a *causal setting* C, we mean a triple of BAT Σ, action sequence z, and goal ϕ, i.e. $C = \langle\Sigma, z, \phi\rangle$ s.t. $\Sigma \models exec(z) \wedge [z]\phi$. When Σ is fixed, we write $C = \langle z, \phi\rangle$ instead.

Define $z' \subseteq z := \exists z'', z = z' \cdot z''$, i.e. z' is a prefix subsequence of z. Given a narrative z, and a prefix z', we consider the counterfactual of z' being absent, a.k.a $Filter(z\backslash z')$, where $Filter(\cdot)$ is recursively defined as: (1) $Filter(\langle\rangle) := \langle\rangle$; (2) for any prefix $z''' \subseteq z\backslash z'$, wlog, assuming $z''' = z^\star \cdot a$, then

$$Filter(z''') := \begin{cases} Filter(z^\star) \cdot a & \Sigma \models [Filter(z^\star)]Poss(a) \\ Filter(z^\star) & otherwise \end{cases}$$

[1] Free variables are implicitly universally quantified from the outside. The $\Box$ modality has lower syntactic precedence than the connectives, and $[\cdot]$ has the highest priority.

Hence $Filter(z\backslash z')$ is the counterfactual (subsequence) of z where the prefix z' is removed and all illegal actions, due to the absence of z', are removed.

Definition 3.2 (Achievement cause for narratives). Given a causal setting $C = \langle z, \phi\rangle$, such that $\Sigma \models \neg\phi$ and $\Sigma \models [z]\phi$. We call a prefix sequence z' of an action sequence z, i.e. $z' \subseteq z$, a cause under C if

(1) $\Sigma \models [z'']\phi$ for all z'' such that $z' \subseteq z'' \subseteq z$;
(2) $\exists z^\star.z^\star = Filter(z\backslash z')$ and $\Sigma \models [z^\star]\neg\phi$;
(3) no subsequence of z' holds for Items 1 and 2.

Item (1) is a *necessary* condition while item (2) is a *sufficient* condition in terms of counterfactual. item (3) is a minimal condition (in the sense of sequence length), namely, we are interested in the minimal prefix of z that satisfies items (1) (2). Intuitively, the subsequence z' is a cause of ϕ under BAT Σ and narrative z, if it is the minimal subsequence that after executing it, ϕ always holds (item (1)) and in the counterfactual that it absents, the remaining legal actions will not change the truth of ϕ (item (2)).

Example 3.3. Consider the BAT Σ in Eq.(1). Suppose the goal is $\phi_1 := Broken(C)$. Clearly, for the narrative $z = pickup(C) \cdot drop(C) \cdot pickup(D)$, we have $\Sigma \models exec(z) \wedge [z]\phi_1$ and the achievement cause here is the prefix $z' = pickup(C) \cdot drop(C)$: in the contingency of z' being absent, the action $pickup(D)$ alone would not cause $Broken(C)$ under Σ. Hence, our notion of cause can successfully identify some redundant actions in a narrative that is irreverent to the achievement of a goal. □

Example 3.4. Let Σ be as above, consider another *disjunctive* goal $\phi_2 := Holding(C) \vee Holding(D)$, for the narrative $z = pickup(C) \cdot pickup(D)$, we have $\Sigma \models exec(z) \wedge [z]\phi_2$ and the achievement cause here is the prefix $z' = pickup(C) \cdot pickup(D)$: in the contingency of z' being absent, $Filter(z\backslash z') = \langle\rangle$, hence under which ϕ_2 would not hold. The action $pickup(C)$ is *not* (but *part of*) a cause, as even if it is absent, the remaining sequence $pickup(D)$ make $Holding(D)$ true, causing the goal ϕ_2 holds. □

Some might claim that this is counter-intuitive as it is ultimately $pickup(C)$ that achieves the goal (no matter if $pickup(D)$ occurs or not, ϕ_2 holds after $pickup(C)$). In fact, this is a limitation of counterfactual cause: counterfactual cause suffers in handling *preemption*. Preemption refers to that two competing events try to achieve the same effect and the latter of these fails to do so, as the earlier event has already achieved the effect. In general, our notion of the cause will disregard the temporal order of occurrences of multiple atomic competing events and include all the competing events as a cause. In an extended report [11], we expand on this discussion.

4 CONCLUSION

In this paper, we propose a notion of cause based on counterfactual. In the context that a narrative leads to a goal, we show that our notion of cause generalizes naturally to a notion of achievement cause. In terms of future work, it is interesting to see how our result can be extended to an epistemic setting, just like [8].

ACKNOWLEDGMENTS

This work has been supported by a Royal Society University Research Fellowship.

REFERENCES

[1] Vitaliy Batusov and Mikhail Soutchanski. 2018. Situation calculus semantics for actual causality. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. AAAI Press, New Orleans, Louisiana, USA, 1744–1752.

[2] Vaishak Belle. 2023. Counterfactual explanations as plans. In *The 39th International Conference on Logic Programming*. Open Publishing Association.

[3] Joseph Y Halpern. 2015. A modification of the Halpern-Pearl definition of causality. In *IJCAI*. ijcai, AAAI Press, Buenos Aires, Argentina, 3022–3033.

[4] Joseph Y Halpern. 2016. *Actual causality*. MiT Press, Cambridge, Massachusetts.

[5] Joseph Y Halpern and Christopher Hitchcock. 2011. Actual causation and the art of modeling. *arXiv preprint arXiv:1106.2652* abs/1106.2652 (2011), 24.

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

[7] David Hume. 1748. *An Enquiry Concerning Human Understanding*. Reprinted by Open Court Press, LaSalle, France.

[8] Shakil M Khan and Yves Lespérance. 2021. Knowing why—on the dynamics of knowledge about actual causes in the situation calculus. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*. ACM, Virtual Event, 701–709.

[9] Gerhard Lakemeyer and Hector J Levesque. 2005. Semantics for a useful fragment of the situation calculus. In *IJCAI*. Professional Book Center, Edinburgh, Scotland, UK, 490–496.

[10] Hector J Levesque, Raymond Reiter, Yves Lespérance, Fangzhen Lin, and Richard B Scherl. 1997. GOLOG: A logic programming language for dynamic domains. *The Journal of Logic Programming* 31, 1-3 (1997), 59–83.

[11] Daxin Liu and Vaishak Belle. 2025. What Is a Counterfactual Cause in Action Theories? arXiv:2501.06857 [cs.AI] https://arxiv.org/abs/2501.06857

[12] Judea Pearl. 2000. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press* 19, 2 (2000), 3.

[13] Judea Pearl. 2009. *Causality*. Cambridge University Press, Cambridge, UK.

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