Explaining Agent Behaviour via Causal Analysis of Mental States (Extended Abstract)

Maryam Rostamigiv and Shakil M. Khan

University of Regina, Regina, Saskatchewan, Canada
{maryam.rostamigiv,shakil.khan}@uregina.ca
http://www.cs.uregina.ca/~skhan

Abstract. This paper extends previous work on rational agents and epistemic causation in the situation calculus to devise an explanatory framework. It incorporates agents' prioritized goals and intentions, utilizes a black-box goal recognition module, and accommodates causal analysis of observed effects involving knowledge and intentions, caused by knowledge-producing and intention-altering actions, respectively. Leveraging an action theory and mental state formalization, it then illustrates –through a theory of mind-grounded model of explanation– that, in contrast to purely machine learning-based systems, knowledge representation-based systems might indeed be more suitable for generating explanations of observed behaviour.

Keywords: Actual Cause · Causal Knowledge · Intentions · Theory of Mind · Explainable Agency · Situation Calculus · Logic.

1 Introduction

In recent years, researchers have become increasingly interested in developing transparent AI systems whose behaviour can be easily understood. To this end, numerous studies have explored how decisions produced by otherwise opaque sub-symbolic approaches can be explained. This has also led to a renewed interest in the study of explainability in knowledge representation (KR), as advocates of KR argue that its declarative nature makes it cognitively more suited for explanation purpose.

Over the years, there has been some work on formalizing explanations in KR [16,12,19,20,14,3]. Motivated by this, in this paper we also investigate the explanatory potential of KR-based systems, although from an entirely novel perspective. In particular, we use causal analysis of mental states to sketch one such system that demonstrates and reinforces that these systems might indeed be more understandable as they allow for commonsensical and intuitive formalization of explanations.

Our framework is based on the situation calculus (SC) [13,17], a model of knowledge [15,18] and intentions [6] in the SC, and a formalization of actual causation [2] and causal knowledge [9] therein. We extend the framework to include

goal change due to *request* communication actions and to support causal analysis of conative effects (i.e. effects involving agent motivation). We also utilize a black-box module for recognizing agents' intentions. Using these, we propose a definition of explanation of agent behaviour relative to observed effects, one that is grounded in theory of mind. Our proposal here is informal, and we mostly focus on an example to illustrate the idea; a full-fledged formal version is available in [10].

2. Actions, Mental States, and their Dynamics

Our base framework for this is the Situation Calculus (SC), which is a second-order (SO) language for modeling and reasoning about dynamic systems where all changes are result of named actions. Here, a possible state of the domain is represented by a situation. The initial situation S_0 denotes the empty sequence of actions and $do(a, s)$ denotes the successor situation to s resulting from performing the action a. Thus the domain of situations can be viewed as a tree, where the root of the tree is the initial situation S_0 and the arcs represent actions. Properties whose truth values vary from situation to situation, are called fluents. 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.

In the SC, a dynamic domain is formalized using an action theory $\mathcal{D}$ that includes the following set of axioms: (1) first-order (FO) action precondition axioms (APA), one per action, (2) (FO) successor-state axioms (SSA), one per fluent, that succinctly encode both effect and frame axioms and specify exactly when the fluent changes, (3) (FO) initial state axioms describing what is true initially, (4) (FO) unique names axioms for actions, and (5) (SO) domain-independent foundational axioms describing the structure of situations [11].

Following [15,18], we model knowledge using a possible worlds account adapted to the SC. There can now be multiple initial situations. Using an accessibility relation K, the knowledge of an agent d is defined as a necessity operator over K. K is constrained to be reflexive and Euclidean (and thus transitive) in the initial situations to capture the fact that the agent's knowledge is true, and that it has positive and negative introspection.

In our framework, the dynamics of knowledge is specified using a SSA for K that supports knowledge expansion as a result of sensing actions as well as "inform" communication actions. As shown in [18], the constraints on K then continue to hold after any sequence of actions since they are preserved by the SSA for K.

Thus to model knowledge, we will use a theory that is similar to $\mathcal{D}$ above, but with modified foundational axioms to allow for multiple initial epistemic states. Also, action preconditions can now include knowledge preconditions and initial state axioms can now include axioms describing the epistemic states of the agents. Finally, the preconditions of *inform* and aforementioned axioms for K are included. See [17] and [5] for details of these.

Following Khan and Lespérance (KL) [8], we will utilize the sort of *paths* in the SC, which are essentially infinite sequences of executable situations. KL

[7] showed how one can interpret arbitrary computational tree logic (CTL*) [4] formulae within SC with paths. Paths are useful for formalizing future-oriented concepts such as goals, intentions, and other motivational states. We assume that our theory $\mathcal{D}$ includes the axiomatization for paths.

In [6], KL proposed a formalization of prioritized goals (p-goals), intentions, and their dynamics in the SC. The account supports a rich specification of the goals of an agent. In their agent theory, an agent can have multiple goals/desires at different priority levels, possibly inconsistent with each other. They assume that goals are totally ordered with respect to the priority ordering. Each goal is specified using its own goal-accessibility relation G, parameterized by the priority level. KL defined intentions/chosen goals, i.e. the goals that the agent is actively pursuing, as the maximal set of highest priority goals that are consistent with each other and with the agent's knowledge; semantically, this is handled by taking a prioritized intersection of goal-accessibility relations. Their model of goals supports the specification of general temporally extended goals (represented by CTL* formulae), not just achievement goals. They also specified how these goals evolve when actions/events occur, when the agent's knowledge changes, or when the agent adopts or drops a goal. This is specified via the SSA for G. Their formalization of prioritized goal dynamics ensures that the agent always optimizes their intentions. They will abandon a chosen goal ϕ if an opportunity to commit to a higher priority goal, that is inconsistent with ϕ, arises. As such their model displays an idealized form of rationality.

We propose to adopt and modify KL's framework to accommodate multiple agents, by adding an agent argument to the hierarchy of goal-accessibility relations. We also modify goal dynamics by introducing a request action $req(d, d', \phi)$, that can be used by an agent d to request another agent d' to adopt a p-goal ϕ, simplifying the model to only include extremely cooperative agents that always adopts the requested goal as their intentions (even if it is inconsistent with their current intentions; note, the requestee's intentions do remain consistent). For this we propose the APA for this *req* action and update the SSA for G; see [10] for the formal details.

3. Causation and Explanation

Given a history of actions/events (often called a scenario) and an observed effect, *actual causation* involves figuring out which of these actions are responsible for bringing about this effect. When the effect is assumed to be false before the execution of the actions in the scenario and true afterwards, the notion is referred to as *achievement (actual) causation*. Based on Batusov and Soutchanski's original proposal [2], KL recently offered a definition of achievement cause in the SC [9]. Both of these frameworks assume that the scenario is a linear sequence of actions, i.e. no concurrent actions are allowed. KL's proposal can deal with epistemic causes and effects; e.g., an agent may analyze the cause of some newly acquired knowledge, and the cause may include some knowledge-producing action, e.g. *inform*. They showed that an agent may or may not know all the causes of an effect, and can even know some causes while not being sure about others.

In this framework, causes are computed relative to a *causal setting* consisting of a domain theory $\mathcal{D}$, a scenario σ (which, again, is a linear sequence of actions), and an effect φ. Since all changes in the SC result from actions, they identified the potential causes with a set of ground action terms occurring in σ. However, since σ might include multiple occurrences of the same action, one also needs to identify the situations where these actions were executed; for brevity, we will ignore this component here. The underlying idea of computing causes is as follows. If some action α of the action sequence in σ triggers the formula φ 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 φ back to false, then α is an actual cause of achieving φ in σ. Moreover, note that α might have been non-executable initially; so other preceding actions that contributed to ensuring that its preconditions are brought about must also be considered as (indirect) cause of φ. Similarly, α might have only brought about φ conditionally, and other preceding actions that achieved those conditions must be considered as (indirect) cause of φ. Using this reasoning, 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.

We propose to extend KL's framework in [9] to include intentions in effect formulae; see [10] for how this can be done. With this extension, we can now analyze the causes of an agent having some intention ψ in some scenario. In our framework, such effects are usually caused by request actions by others.

To propose a model of explanation on top of this framework, we need one last component, which for this work is considered to be a black-box module. Given an agent d, an action α, and a scenario s, the component under consideration is a goal recognition module, which recognizes the intention of d behind performing α in s. With this, we are now ready to give a definition of explanation. Just like causes, explanations in our framework are simply actions from the scenario. However, as we will see, they are not simply causes.

Explanation We say that the behaviour of a group of agents captured by a scenario s relative to the observation that φ can be explained by the action a if and only if a is a cause of φ in s; or a causes the intention behind some explanation of φ in s, i.e. some other action a' explains φ in s, the agent of a' is d', d' is recognized to have the intention that ψ behind performing a' in s, and a was the cause of this intention in s'.

4. Reasoning Example

We consider a domain where we have two rescue drone agents, D_1 and D_2, navigating through four locations, L_s, L_d, L_1, and L'_1, and managed by a controller agent D_c. We have a non-fluent relation, $\mathit{Route}(l, l')$, that represents a flight path from location l to l'. The routes are defined as: from L_s to L_1, from L_s to L'_1, from L_1 to L_d, and from L'_1 to L_d. In this domain, there are four self-explanatory actions (given d and l are drones and locations, resp.): $\mathit{takeOff}(d, l)$, $\mathit{flyTo}(d, l, l')$, $\mathit{land}(d, l)$, and the aforementioned communicative action *inform*.

The fluents in this domain are $At(d, l, s)$, $Flying(d, s)$, $Vis(d, l, s)$ (i.e. d has visited l in s), and $TStorm(l, s)$ (i.e. there is a thunderstorm at l in s).

The preconditions for the above actions are as follows. Agent d can takeoff at location l in situation s iff it is at l in s and it is not flying in s. d can fly to l' from l in s iff it is at l in s, it is flying in s, there is a route from l to l', and d does not know that there is a thunderstorm at l' in s. d can land at l in s iff it is at l in s and it is flying in s. Finally, d can inform d' that Φ in s iff d knows in s that Φ and it does not know in s that d' knows that Φ.

The SSA for the above fluents are as follows. d is at location l after executing action a in situation s iff a refers to d's action of flying from some location l' to l, or d was already at l in s and a is not its action of flying to a different location l'. d is flying in $do(a, s)$ iff a is d's action of taking off at some location l, or it was already flying in s and a is not its action of landing at some location. d has visited l in $do(a, s)$ iff a refers to its action of flying to l from some other location, or it has already visited l in s. Finally, there is a thunderstorm at l in $do(a, s)$ iff this is the case in s (for simplicity, we treat this as a non-fluent).

The Knowledge of agents initially are as follows. Drone D_1 knows that it is at location L_s, that it is not flying, and that it has only visited L_s. Moreover, it does not know that there is a storm at location L_1, but knows that there are no storms at L'_1 and L_d. There is indeed a thunderstorm at location L_1 and the controller agent D_c knows this. Finally, D_c does not know however that the other agents know this fact.

Assume that, initially our drone agent D_1 has the following two p-goals: $\phi_0 = \Diamond At(D_1, L_d)$, i.e. that it is eventually at L_d, at the highest priority level, and $\phi_1 = Vis(D_1, L_1) \mathcal{B} Vis(D_1, L_d)$, i.e. that it visits L_1 before it visits L_d, at a lower priority level, respectively. Also, D_c does not have any initial p-goals.

To see an example of intention dynamics, note that in our example, we can show that the agent D_1 will have the intention that $\Diamond Vis(D_1, L'_1)$ after D_1 takes off from L_s, D_c informs D_1 that there is a thunderstorm at L_1, and D_c requests D_1 to eventually visit L'_1, starting in the initial situation; thus we can show that D_1 intends to eventually visit L'_1 afterwards. But D_1 will not have the intention that ϕ_1 afterwards as it has become impossible for D_1 to visit L_1 due to its knowledge of the thunderstorm at L_1.

Next, let us consider an example of causation relative to conative effects. Assume that $\sigma = do([takeOff(D_1, L_s), inform(D_c, D_1, TStorm(L_1)), req(D_c, D_1, \Diamond Vis(D_1, L'_1)), \quad inform(D_c, D_2, TStorm(L_1)), \quad req(D_c, D_2, \Diamond Vis(D_1, L'_1)), flyTo(D_1, L_s, L'_1), flyTo(D_1, L'_1, L_d)], \quad S_0)$. There are 7 actions in this scenario. For convenience, we will use $\vec{\alpha_i}$ to denote the first i actions in this trace, and so $do([\vec{\alpha_5}], S_0)$ is the situation obtained from executing the first 5 actions starting in S_0. Now assume that we want to reason about the causes of the effect $\varphi_1 = Int(D_1, \Diamond Vis(D_1, L'_1))$ in scenario $\sigma_1 = do([\vec{\alpha_5}], S_0)$. We can show that D_c's request to D_1 to eventually visit L'_1 is the only cause of D_1's intention that $\Diamond Vis(D_1, L'_1)$ in σ_1. Thus, e.g., $req(D_c, D_2, \Diamond Vis(D_1, L'_1))$ is not a cause.

Finally, we would like to explain the behaviour of drones as modeled by situation/scenario σ above relative to the effect that $\varphi_2 = Vis(D_1, L'_1)$, i.e. we want to

understand why D_1 visited L_1' (rather than the usual path of L_1). As expected, we can show that agent behaviour in σ w.r.t. visiting L_1' can be explained by D_1's action $\mathit{flyTo}(D_1, L_s, L_1')$. But perhaps more interestingly, assuming that the intention behind D_1's action of $\mathit{flyTo}(D_1, L_s, L_1')$ in σ was recognized to be $\Diamond \mathit{Vis}(D_1, L_1')$, we can further explain agent behaviour via the causes of having this intention. This will in turn reveal that D_1 had this intention due to D_c's request to D_1 to eventually visit L_1', and thus agent behaviour w.r.t. D_1 visiting L_1' can be explained by this request action as well.

5. Discussion and Conclusion

In this paper, we sketched an account of causal reasoning about motivations. Using this, we offered a novel take on explainable AI that is grounded in theory of mind: agent behaviour in our framework can be explained via the causal analysis of observed effects, which in turn can trigger the analysis of their mental states.

As mentioned, there has been some work on formalizing explanation in KR. For instance, in his early work, Shanahan [19] proposed a deductive and an abductive approach to explanation in the situation calculus, both of which are based on default reasoning. More recently, Shvo et al. [20] proposed a belief revision-based account of explanation. In their framework, a formula ϕ explains another formula ψ if revising by ϕ makes the agent believe ψ and the agent's beliefs are still consistent afterwards. In [3], Dennis and Oren used dialogue between the user and a Belief-Desire-Intention (BDI) agent system to explain why the agent has chosen a particular action. Their approach aims to identify any divergence of views that exist between the user and the BDI agent relative to the latter's behaviour and allows for an interactive and user-friendly explanation process. Miller [14] proposed a contrastive explanation model based on structural causal models to enhance understanding and trust in AI decision-making. Finally, Sridharan et al. [21,22] proposed an explainable robotic architecture by integrating step-wise refinement, non-monotonic reasoning, probabilistic planning, and interactive learning. However, none of these proposals perform causal analysis of agent motivation or employ such reasoning for explaining agent behaviour. In fact to the best of our knowledge, our proposal is the first and the only attempt to this end.

Our current formalization is limited in many ways. For instance, we only allow deterministic and fully observable actions. Scenarios in our framework are linear, i.e. we assume that the order of action occurrence is known. When dealing with causation and explanations, we computed achievement causes only. Incorporating other types of causes, e.g. maintenance causes [1], would have allowed us to explain effects further and in finer details. We leave these for future work.

Acknowledgments

We thank the anonymous reviewers of CAKR 2023, XLoKR 2023, and ECAI 2023 for directing us to the relevant literature.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2022-03433].

Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [numéro de référence RGPIN-2022-03433].

References

1. Batusov, V., Soutchanski, M.: Situation calculus semantics for actual causality. In: Gordon, A.S., Miller, R., Turán, G. (eds.) Proceedings of the Thirteenth International Symposium on Commonsense Reasoning, COMMONSENSE 2017, London, UK, November 6-8, 2017. CEUR Workshop Proceedings, vol. 2052. CEUR-WS.org (2017)
2. Batusov, V., Soutchanski, M.: Situation calculus semantics for actual causality. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 1744–1752. AAAI Press (2018)
3. Dennis, L.A., Oren, N.: Explaining BDI agent behaviour through dialogue. Auton. Agents Multi Agent Syst. **36**(1), 29 (2022)
4. Emerson, E.A., Halpern, J.Y.: "sometimes" and "not never" revisited: On branching versus linear time temporal logic. J. ACM **33**(1), 151–178 (1986)
5. Khan, S.M., Lespérance, Y.: ECASL: a model of rational agency for communicating agents. In: Dignum, F., Dignum, V., Koenig, S., Kraus, S., Singh, M.P., Wooldridge, M.J. (eds.) 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25-29, 2005, Utrecht, The Netherlands. pp. 762–769. ACM (2005)
6. Khan, S.M., Lespérance, Y.: A logical framework for prioritized goal change. In: van der Hoek, W., Kaminka, G.A., Lespérance, Y., Luck, M., Sen, S. (eds.) 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), Toronto, Canada, May 10-14, 2010, Volume 1-3. pp. 283–290. IFAAMAS (2010)
7. Khan, S.M., Lespérance, Y.: Infinite paths in the situation calculus. Tech. Rep. EECS-2015-05, Department of Electrical Engineering and Computer Science, York University, Toronto, Canada (2015)
8. Khan, S.M., Lespérance, Y.: Infinite paths in the situation calculus: Axiomatization and properties. In: Baral, C., Delgrande, J.P., Wolter, F. (eds.) Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016. pp. 565–568. AAAI Press (2016)
9. Khan, S.M., Lespérance, Y.: Knowing why - on the dynamics of knowledge about actual causes in the situation calculus. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021. pp. 701–709. ACM (2021)
10. Khan, S.M., Rostamigiv, M.: On explaining agent behaviour via root cause analysis: A formal account grounded in theory of mind. In: Proceedings of the 26th

European Conference on Artificial Intelligence ECAI, 30.09 - 5.10, 2023, Kraków, Poland (2023)
11. Levesque, H.J., Pirri, F., Reiter, R.: Foundations for the situation calculus. Electronic Transactions on Artificial Intelligence (ETAI) **2**, 159–178 (1998)
12. Lifschitz, V., Rabinov, A.: Miracles in formal theories of action. Artif. Intell. **38**(2), 225–237 (1989)
13. McCarthy, J., Hayes, P.J.: Some philosophical problems from the standpoint of artificial intelligence. Machine Intelligence **4**, 463–502 (1969)
14. Miller, T.: Contrastive explanation: a structural-model approach. Knowl. Eng. Rev. **36**, e14 (2021)
15. Moore, R.C.: A formal theory of knowledge and action. In: Formal Theories of the Commonsense World. pp. 319–358. Ablex (1985)
16. Morgenstern, L., Stein, L.A.: Why things go wrong: A formal theory of causal reasoning. In: Shrobe, H.E., Mitchell, T.M., Smith, R.G. (eds.) Proceedings of the 7th National Conference on Artificial Intelligence, St. Paul, MN, USA, August 21-26, 1988. pp. 518–523. AAAI Press / The MIT Press (1988)
17. Reiter, R.: Knowledge in Action. Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press, Cambridge, MA, USA (2001)
18. Scherl, R.B., Levesque, H.J.: Knowledge, action, and the frame problem. Artificial Intelligence **144**(1-2), 1–39 (2003)
19. Shanahan, M.: Explanation in the situation calculus. In: Bajcsy, R. (ed.) Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993. pp. 160–165. Morgan Kaufmann (1993)
20. Shvo, M., Klassen, T.Q., McIlraith, S.A.: Towards the role of theory of mind in explanation. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems - Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9-13, 2020, Revised Selected Papers. Lecture Notes in Computer Science, vol. 12175, pp. 75–93. Springer (2020)
21. Sridharan, M.: REBA-KRL: refinement-based architecture for knowledge representation, explainable reasoning and interactive learning in robotics. In: Giacomo, G.D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., Lang, J. (eds.) ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2935–2936. IOS Press (2020)
22. Sridharan, M., Meadows, B.: Towards a theory of explanations for human-robot collaboration. Künstliche Intell. **33**(4), 331–342 (2019)