Towards a Logical Account for Human-Aware Explanation Generation in Model Reconciliation Problems

Doctoral Consortium

Stylianos Loukas Vasileiou
Washington University in St. Louis
St. Louis, USA
v.stylianos@wustl.edu

ABSTRACT

A model reconciliation problem focuses on producing explanations for human users who have varying expectations of the AI agent. This research explores the development of a general framework for generating human-aware explanations in such problems. We face two primary challenges: creating an expressive and efficient framework that generates personalized and persuasive explanations for users, and interactively incorporating users' knowledge, beliefs, and preferences into the explanation process. We propose that a logic-based framework is well-suited for identifying, representing, and offering robust and tailored explanations to human users in model reconciliation scenarios.

KEYWORDS

Human-Aware AI; Explainable AI Planning; Knowledge Representation & Reasoning

ACM Reference Format:
Stylianos Loukas Vasileiou. 2023. Towards a Logical Account for Human-Aware Explanation Generation in Model Reconciliation Problems: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 OVERVIEW

The increasing ubiquity of AI systems in our everyday life necessitates a need for creating systems that explain their decisions in order to assure their human users that their reasoning is correct and their conclusions are sound. Toward this end, Explainable AI Planning (XAIP) [3] has garnered a lot of interest due to its focus on helping human users interface with AI systems in sequential decision-making problems. On the other hand, understanding the psychological theories behind human-human interactions will hew us closer to the birth of human-aware AI systems that collaborate and interact with humans in a naturalistic way [4]. For example, according to the Theory of Mind [8],[1] when two humans collaborate on a task, they usually develop approximate models of the goals and capabilities of each other, and use them to support a fluid team performance. It is only a small step to suggest that AI agents interacting with humans will also need to take this implicit mental modeling into account.

[1]The theory of mind is the ability to attribute mental states (beliefs, intents, knowledge, etc.) to others and recognize that these mental states may differ from one's own.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Indeed, this is the view of the model reconciliation problem (MRP) [2] – a popular paradigm within XAIP that considers AI agents that exhibit explainable behavior. To synthesize explainable behavior, AI agents interacting with human users go beyond planning with their own models of the world and incorporate the (mental) models of the human users in their deliberative processes. As such, explanations in this context are couched in terms of features from the agents' and users' models. While MRP has gathered a lot of traction due to its "human-like" explanation generation process, many advancements towards that end are usually based on some simplistic assumptions [1, 9, 10]. The most common assumption is that the user's mental model is available to the agent a-priori. Moreover, most existing approaches consider one-shot explanation approaches, hence failing to exploit the potential advantage of an interdependent multi-shot explanation generation process. For example, in a multi-shot approach a user would be able to interact, in a collaborative manner, with the agent's explanation process and provide information about their knowledge, beliefs or preferences. Therefore, through that process the agent would be able to learn a more accurate representation of the user's model, which can not only lead to generating personalized explanations that enhance the user's understanding, but also improve (or correct) the agent's model if it is clashing with the ground truth. Finally, explanations generated by most MRP methods are grounded in the syntax of planning, thus failing to generalize to problems beyond planning.

All the above give rise to an opportunity for developing an account that reasons efficiently about decision-making tasks and provides explanations to human users in an interactive and easy-to-understand fashion. Therefore, in this research, we propose to investigate the creation of a general account for human-aware AI agents exhibiting explainable behavior that advances model reconciliation. The key challenges in this research are two-fold: (1) How to create an expressive and efficient account that generates robust and personalized explanations for human users; and (2) How to incorporate the users' knowledge, beliefs, and preferences in the agent's explanation process in an interactive fashion. To address these challenges, we propose the following objectives:

- **Objective 1:** Formalize a logic-based account for MRP that is sufficiently expressive and general.
- **Objective 2:** Develop efficient algorithms that find robust and personalized explanations.
- **Objective 3:** Develop methods for an interactive multi-shot explanation process.

In summary, we hypothesize that *a logic-based account is well equipped to identify, represent, and provide robust and personalized*

explanations through interactions between an agent and a human user for (sequential) decision-making problems, so long as these problems can be expressed in a logical formalism. We are optimistic that the efforts of this research will help push us a step closer towards a trustworthy, transparent, and accountable AI-driven society.

2 LOGIC-BASED MODEL RECONCILIATION

The principal objective of this research is to characterize MRP from a logic-based perspective that: (i) accounts for discrepancies and uncertainties in the agent's and user's models; (ii) is sufficiently general and expressive; and (iii) can produce efficient and scalable solutions. Indeed, in [12] we introduced the following formulation, where we use $\models_L$ to represent logical entailment in a logic L:

> [**L-MRP**] Given two knowledge based KB_a and KB_h and a formula φ in a logic L, where $KB_a \models_L \varphi$ and $KB_h \not\models_L \varphi$, the goal is to identify an explanation $\epsilon = \langle \epsilon^+, \epsilon^- \rangle$, $\epsilon^+ \subset KB_a$ and $\epsilon^- \subset KB_h$, s.t. when it is used to *update* KB_h to $\widehat{KB_h}^{\epsilon}$, then $\widehat{KB_h}^{\epsilon} \models_L \varphi$.

In this setting, KB_a represents the knowledge base (e.g., model) of the agent, KB_h represents the agent's knowledge base of the human user, and φ is the target of the explanation process (e.g., explanandum). One challenge here is specifying how KB_h is updated by an explanation $\epsilon = \langle \epsilon^+, \epsilon^- \rangle$. A trivial and intuitive update process is $KB_h = (KB_h \cup \epsilon^+) \setminus \epsilon^-$, where new formulae ϵ^+ from the agent are added and erroneous formulae ϵ^- from the user are removed to ensure consistency. Note that the explanation generation problem is solved when an explanation ϵ for φ is found and provided to the user. Nevertheless, if the agent has an inaccurate assumption of KB_h, the user might fail to understand the explanation (we will revisit this limitation in Objective 3).

Besides finding an explanation ϵ, the agent also needs to present that explanation to the user. Given knowledge bases KB_a and KB_h and an explanandum φ, there might be several explanations for φ from KB_a for KB_h. As such, an agent might prefer an explanation that requires the least amount of effort in presenting it to the user. One way to characterize the effort of the agent when presenting an explanation is to associate a cost to the elements of the explanation.

2.1 Computing Explanations

Computing explanations for the explanandum φ from KB_a for KB_h should be done in a robust and efficient manner. From the perspective of planning, an algorithm developed for MRP should incorporate some of its core assumptions: (1) KB_a and KB_h should encode the version of the same problem (e.g., planning) of the agent and human user, respectively; (2) KB_a is correct and complete, and only KB_h can contain errors or omissions; and (3) KB_a appropriately entails the explanandum φ. The first assumption is reasonable and follows closely the definition of a model reconciliation planning problem. The last two assumptions stem from fact that the explaining agent bases its explanations on the view (or model) of the specific problem [7]. Therefore, the agent should believe that its model is correct and complete, and that its model appropriately entails the explanandum. Together, these three assumptions imply that each erroneous formula in KB_h will have a corresponding correct formula in KB_h. One can see that this is a property that can be exploited when searching for explanations. Therefore, the primary challenge in this objective is to identify and leverage various pruning heuristics and search strategies that can speed up the explanation search process.

In our earlier work [12], we empirically demonstrated that our logical MRP account can not only be applied to classical planning problems but also hybrid systems planning problems with durative actions, processes, and events. Notably, our account advanced the state-of-the-art in model reconciliation for explanation generation within XAIP along two key dimensions: (1) it improved the scalability for some types of classical planning problems; and (2) generalized the model reconciliation approach to other types of planning problems beyond classical planning.

Further, in [11] we developed an algorithm for problems beyond the realm of planning. Specifically, we exploited the notion of *minimal unsatisfiable sets* (MUSes) [5] for generating explanations[2] and empirically demonstrated its efficacy and generality on a plethora of instances. Overall, this approach has the general advantage of being able to deal with problems coming from different settings, so long as these problems can be encoded into a logical formalism for which satisfiability of subsets can be decided.

3 ONGOING AND FUTURE WORK

A key underlying assumption thus far has been that the knowledge base of the agent KB_a represents the ground truth and that the knowledge base of the human user KB_h is known to the agent a-priori. As this fails to exploit the potential advantage of an interconnected explanation generation process. We plan to investigate a dialogical extension, where the agent engages in dialogue with the human user and, through that process, the agent is able to learn more accurate representations of the user's model, as well as correct its own model if it is clashing with the ground truth (Objective 3). To enable a dialogical approach, we will consider the machinery of argumentation frameworks. This is motivated by the fact that humans reason and understand better when they engage in an argumentative process [6]. Two main challenges here are defining the underlying semantics of an argumentative account, and defining how to interactively update the knowledge bases.

Moreover, we are interested in extending our logic-based MRP formulation to problems that are probabilistic in nature. For instance, we can consider probabilistic knowledge bases of the agent and the user, where each knowledge base consists of a set of formulae representing a specific problem instance and each formula f has an associated weight $p(f)$ that corresponds to the probability of f being true. A probabilistic knowledge base provides a way of representing various uncertainties in the world, and can therefore be used to generate the *most-probable explanations* by using state-of-the-art probabilistic reasoning methods. Finally, we are also interested in pursuing a more personalized approach to generating explanations. Naturally, in human-to-human explanations, the explainer will usually decide how much detail to include in the explanation by choosing a conceptual model that they think will mesh with that of the explainee. Thus, explanations can be given at different levels of abstraction, based on different conceptual models.

[2] MUSes are typically a subset of formulae from an unsatisfiable knowledge base that "explain" why the knowledge base is unsatisfiable.

REFERENCES

[1] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2019. Balancing Explicability and Explanations in Human-Aware Planning. In *IJCAI*. 1335–1343.

[2] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*. 156–163.

[3] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable Planning. *CoRR* abs/1709.10256 (2017).

[4] Subbarao Kambhampati. 2019. Synthesizing Explainable Behavior for Human-AI Collaboration. In *Proceedings of AAMAS*. 1–2.

[5] Mark H Liffiton and Karem A Sakallah. 2008. Algorithms for computing minimal unsatisfiable subsets of constraints. *Journal of Automated Reasoning* 40, 1 (2008), 1–33.

[6] Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* (2011), 57–74.

[7] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[8] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[9] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2018. Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation. In *ICAPS*. 518–526.

[10] Sarath Sreedharan, Alberto Olmo Hernandez, Aditya Prasad Mishra, and Subbarao Kambhampati. 2019. Model-Free Model Reconciliation. In *IJCAI*. 587–594.

[11] Stylianos Loukas Vasileiou, Alessandro Previti, and William Yeoh. 2021. On Exploiting Hitting Sets for Model Reconciliation. In *AAAI*.

[12] Stylianos Loukas Vasileiou, William Yeoh, Tran Cao Son, Ashwin Kumar, Michael Cashmore, and Daniele Magazzeni. 2022. A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems. *Journal of Artificial Intelligence Research* 73 (2022), 1473–1534.