Towards Understanding User Preferences for Explanation Types in Model Reconciliation

Zahra Zahedi[†]*, Alberto Olmo[†]*, Tathagata Chakraborti[‡], Sarath Sreedharan[†] and Subbarao Kambhampati[†]
[†]Department of Computer Science, Arizona State University, Tempe, AZ
Email: { zzahedi, aolmoher, ssreedh3, rao } @ asu.edu
[‡]AI Composition Lab, IBM Research AI, Cambridge, MA
Email: tathagata.chakraborti1@ibm.com

***Abstract*—Recent work has formalized the explanation process in the context of automated planning as one of *model reconciliation* – i.e. a process by which the planning agent can bring the explainee's (possibly faulty) model of a planning problem closer to its understanding of the ground truth until both agree that its plan is the best possible. The content of explanations can thus range from misunderstandings about the agent's beliefs (state), desires (goals) and capabilities (action model). Though existing literature has considered different kinds of these model differences to be equivalent, literature on the explanations in social sciences has suggested that explanations with similar logical properties may often be perceived differently by humans. In this brief report, we explore to what extent humans attribute importance to different kinds of model differences that have been traditionally considered equivalent in the model reconciliation setting. Our results suggest that people prefer the explanations which are related to the effects of actions.**

I. Introduction

Informally, a classical planning problem is defined [1] in terms of a start state, a goal state and a set of operators each of which contain a set of conditions (preconditions) that must be satisfied for that operator to be applicable in state and another set of conditions (effects) that are produced as a result of their application. The solution to a planning problem is therefore a sequence of actions or plan that transforms the start state to the goal state. Work on explanations of plans thus begin with the following question from a human in the loop:

Q. Why this plan?

The root cause of such a question is when the model of the planning problem that the human has (i.e. the human mental model) differs from that of the planner. This means that even though the planner has come up with the best plan in its own model, that plan is no longer the best one when evaluated in the human mental model. Apart from the components of the classical planning model [2] as described above, this is, of course, contingent on the criterion (e.g. optimality) that a plan is being evaluated on and subject to computational capability of the human in the loop. In such scenarios, the process of explanation of a plan becomes one of "model reconciliation" where the planning agent attempts to attain common ground with their model and the human mental model.

This research is supported in part by the AFOSR grant FA9550-18-1-0067, the ONR grants N00014161-2892, N00014-13-1-0176, N00014- 13-1-0519, N00014-15-1-2027, and the NASA grant NNX17AD06G.

* Authors marked with asterix contributed equally.

A. Explanations as Model Reconciliation

In existing work [3], authors have explored how this model reconciliation process unfolds in the classical planning setting, while accounting for the mental model of the human in the loop. Specifically, an explanation of a plan in this framework satisfies conditions (1) and (2) below:

1. An explanation is an update to the human mental model
2. such that there is no better plan in the updated mental model than the given plan.

Clearly, from the perspective of the model of a planning problem, (1) can be in terms of one or more of –

- beliefs of the agent about the current state (as opposed to what the human may be aware of);
- their actual desires or goals (as opposed to that ascribed to it by the human);
- preconditions and effects of actions (as opposed to its capabilities known to the human).

3. **Minimally Complete Explanation (MCE)** is the shortest model explanation that satisfies (1) and (2).

In addition to MCEs, authors in [3] investigate the many forms of model updates and their unique properties as it pertains to the plan explanation process. Authors in [4] explore the effectiveness of this model reconciliation process as a viable means of plan explanations.

Contrastiveness and Selectiveness of Explanations

In recent work [5], authors outline the salient features of an explanation as studied in the literature in the social sciences on explanations in human-human interactions. Three features particularly emphasized are –

- Explanations are *social* – they account for human expectations, in this case, the human mental model;
- Explanations are *contrastive* [6] – they are able to contrast between the explanandum (i.e. the plan being explained) and its foils (other alternative plans); and
- Explanations are *selective* [7] – i.e. explanations are subject to heuristics to decide among many possible, equally valid, explanations.

MCEs (as introduced above) satisfy the contrastive property by ensuring that the plan is better than all possible foils – i.e. in answer to the question *Why this plan (as opposed to all other plans)?*; while being selective in the sense that it chooses the smallest number of conditions or model differences that can answer this question. However, MCEs are known to be non-unique [3] and, as we discussed before, there may be many different kinds of model updates (1) that can satisfy (2) as well as (3). Thus, from the point of view of the selective property of explanations, this poses an interesting challenge for the planner in figuring out how to prioritize different aspects of its model during the explanation process. In the next section, we explore in a preliminary study to what extent such preferences exist in plan explanations as model reconciliation.

II. Study

The following is a preliminary exploration on user preferences over different kinds of model differences based on a typical – logistics [8] – planning domain. In the experiment, we considered a logistics company which has deployed an AI system to assist in their decision-making. This system can automatically collect information about different cities, routes and vehicles to plan the best way for transporting packages. Each user, based on the given information, is asked to make a plan and then evaluate the plan the system has come up with. The given information determines the human mental model which may be different than the system's model. The system provides different kinds of MCEs to convince the user about its plan. The questionnaire is accessible at https://goo.gl/bD8Z1p. Each example had different types of explanations:

- E1 has three different types of explanations which deal with initial conditions, effects and preconditions;
- E2 has two types of explanations that deal with preconditions and goals; and
- E3 also has three types of explanations related to initial conditions, effects and goals.

The experiment was conducted on 38 volunteers (none of them were planning experts). We only studied those who could properly perform the planning task according to the given information. Figure 1 shows the results on each example and people preferences regarding the different types of explanations. From the results, people prefer the types of explanations which are related to the effects of an action in this particular domain. Based on the results in E2, when comparing preconditions and goals, people seemed to prefer preconditions over goals. However, the percentage of people who don't have any preference on these two types in E2 was also 50%. Interestingly, in the original treatment of MCEs [3], [4] *all these explanations were considered equivalent*.

III. Recommendations

In this paper, we have explored the preferences of a human in the loop towards different types of explanations provided by an AI planner in the paradigm of explanations as model reconciliation. Our results suggests that humans prefer explanations that address misunderstandings about the effects of

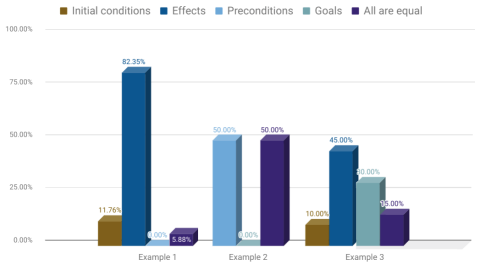

Fig. 1. Percentage of participants who showed preference on effects, preconditions, initial conditions and goal (or none) in the 3 examples.

the agent's actions. This indicates that model reconciliation approaches should do well to incorporate such considerations in the explanation generation process.

- Algorithmically, the easiest way to bias the model reconciliation process towards preferred model updates is to penalize or incentivize certain model updates over others (as evident from the user feedback) during the model space search [3] – i.e. by allowing non-unit costs on model change actions. However, more sophisticated approaches may be required to capture relative importances of the model updates which may, for example, be dependent variables. As such, in contrast to the contrastive property of explanations, their selective property – as explored in [5] – is yet to be explored satisfactorily in the literature on explanations (particularly in the framework of explanations as model reconciliation) and is likely to provide a rich set of research problems going forward.
- Authors in [9] explore the possibility of engaging the explainee in dialog when the latter's model is not known with certainty. Such techniques can be of potential use in figuring out latent preferences of the explainee. In general, these preferences (explored in this report in the context of a typical planning domain) are likely to be domain dependent, as well as specific to the particular human in the loop, and thus must be learned in a given setting in the course of interactions.

References

[1] S. Russell and P. Norvig, *Artificial Intelligence – A Modern Approach*. Prentice Hall, 2003.

[2] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins, "PDDL – The Planning Domain Definition Language," 1998.

[3] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy," in *IJCAI*, 2017.

[4] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan Explanations as Model Reconciliation – An Empirical Study," *HRI*, 2019.

[5] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *AIJ*, 2018.

[6] T. Lombrozo, "Explanation and Abductive Inference," *Oxford Handbook of Thinking and Reasoning*, pp. 260–276, 2012.

[7] D. Hilton, "Social attribution and explanation," in *The Oxford Handbook of Causal Reasoning*.

[8] International Planning Competition, "IPC Competition Domains," https://goo.gl/3yyDn4, 2011.

[9] S. Sreedharan, T. Chakraborti, and S. Kambhampati, "Handling Model Uncertainty and Multiplicity in Explanations as Model Reconciliation," in *ICAPS*, 2018.