Plan Explanations as Model Reconciliation
– *An Empirical Study* –

Tathagata Chakraborti
AI Composition Lab
IBM Research AI
Cambridge, MA, USA
tathagata.chakraborti1@ibm.com

Sarath Sreedharan
Computer Science
Arizona State University
Tempe, AZ, USA
ssreedh3@asu.edu

Sachin Grover
Computer Science
Arizona State University
Tempe, AZ, USA
sgrover6@asu.edu

Subbarao Kambhampati
Computer Science
Arizona State University
Tempe, AZ, USA
rao@asu.edu

***Abstract*—Recent work in explanation generation for decision making agents has looked at how unexplained behavior of autonomous systems can be understood in terms of differences in the model of the system and the human's understanding of the same, and how the explanation process as a result of this mismatch can be then seen as a process of reconciliation of these models. Existing algorithms in such settings, while having been built on contrastive, selective and social properties of explanations as studied extensively in the psychology literature, have not, to the best of our knowledge, been evaluated in settings with actual humans in the loop. As such, the applicability of such explanations to human-AI and human-robot interactions remains suspect. In this paper, we set out to evaluate these explanation generation algorithms in a series of studies in a mock search and rescue scenario with an internal semi-autonomous robot and an external human commander. During that process, we hope to demonstrate to what extent the properties of these algorithms hold as they are evaluated by humans.**

***Index Terms*—Explainable AI, planning and decision-making, human-robot interaction, explanations as model reconciliation.**

I. Introduction

The issue of explanations for AI systems operating alongside or with humans in the loop has been a topic of considerable interest of late [1], [2], especially as more and more AI-enabled components get deployed into hitherto human-only workflows. The ability to generate explanations holds the key [3], [4] towards acceptance of AI-based systems in collaborations with humans. Indeed, in many cases, this may even be *required* by law [5].

Of course, the answer to what constitutes a valid, or even useful, explanation largely depends on the type of AI-algorithm in question. Recent works [7]–[9] have attempted to address that question in the context of human-robot interactions [10] by formulating the process of explaining the decisions of an autonomous agent as a *model reconciliation process* whereby the agent tries to bring the human in the loop to a shared understanding of the current situation so as to explain its decisions in that updated model. This is illustrated in Figure 1a. While these techniques have been developed on theories in the psychology literature [11], [12] built on extensive studies in how humans explain behavior, none of these algorithms have, to the best of our knowledge, been evaluated yet with humans in the loop. As such, it remains unclear whether the theoretical guarantees provided by explanations generated by such algorithms do, in fact, bear out during interactions with humans.

The aim of this paper is then to provide an empirical study of the "explanation as model reconciliation" process, especially as it relates to a human-robot dyad in a mock up version of a typical search and rescue scenario (Section IV) which has been used extensively as an illustrative scenario in existing literature [8], [9]. But before we go there, we provide a brief overview of explanations (Section II) in the planning community and a glossary of relevant terms (Section III).

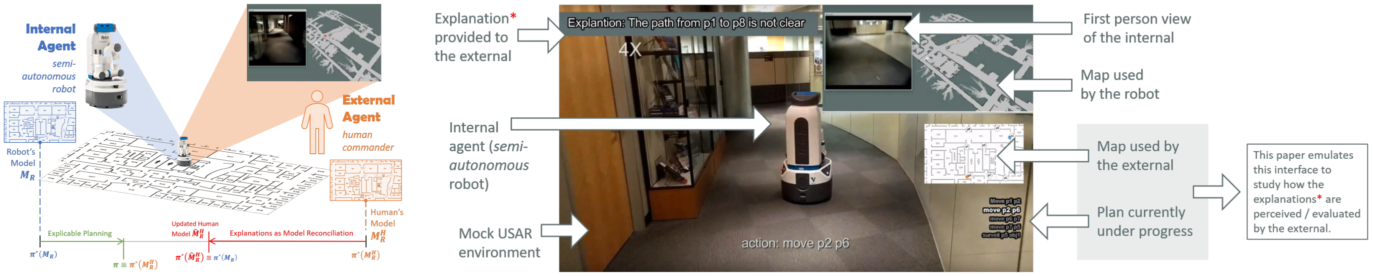
II. A Brief History of Explainable Planning

From the perspective of planning and decision making, the notion of explanations was first explored extensively in the context of *expert systems* [13]. Similar techniques have been looked at for explanations in case based planning systems [14], [15] and in interactive planning [16] where the planner is mostly concerned with establishing the correctness [17] and quality [18], [19] of a given plan *with respect to its own model*. These explanation generation techniques served more as a debugging system for an expert user rather than explanations for situations generally encountered in everyday interactions, which may be referred to as *"everyday explanations"* [20]. A key difference here is that the former is mostly algorithm dependent and explains the *how* of the decision making process whereas the latter is model-based and algorithm-independent and thus explain the *why* of a particular decision in terms of the knowledge that engendered it.

In [21] authors argued that, in a classic case of "inmates running the asylum", most of the existing literature on explanation generation techniques for AI systems are based on the developer's intuitions rather than any principled understanding of the normative explanation process in interactions among humans as has been studied extensively in the fields of philosophy, cognitive science, and social psychology. The authors note that the latter can be a valuable resource for the design of explanation generation techniques in AI systems as well.

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.

Most of the work was done while the first author was at Arizona State University (Computer Science Department) in Tempe, AZ, USA.


Internal
Agent
semi-
autonomous
robot
Robot's
Model
M_R
External
Agent
human
commander
Human's
Model
M_R^H
Explicable Planning
Updated Human
Model $\widehat{M}_R^H$
Explanations as Model Reconciliation
$\pi^*(M_R)$
$\pi \equiv \pi^*(M_R^H)$
$\pi^*(\widehat{M}_R^H) \equiv \pi^*(M_R)$
$\pi^*(M_R^H)$
Explanation* provided to the external
Explantion: The path from p1 to p8 is not clear
4X
Internal agent (semi-autonomous robot)
Mock USAR environment
action: move p2 p6
First person view of the internal
Map used by the robot
Map used by the external
Plan currently under progress
Move p1 p2
move p2 p6
move p6 p7
move p7 p5
surveil p5 obj1
This paper emulates this interface to study how the explanations* are perceived / evaluated by the external.

(a) (b)

Fig. 1: Illustration of a typical [6] urban search and reconnaissance (USAR) scenario with an internal semi-autonomous robot and an external human supervisor. The supervisor has restricted access to a changing environment – thus models (e.g. map of the environment) may diverge in the course of the operation due to the disaster. In this situation, the robot can either choose to generate *explicable plans* by conforming to the expectations of the human or *explain* its plans in terms of their model differences via the *model reconciliation process* (as illustrated in inset 1a). We make use of this setting to study the properties of "model-reconciliation" explanations in a mock interface (Figure 5) to the robot (derived from the interface in inset 1b).

The authors in [20] state the three most important properties of explanations (as accounted for in the existing literature in the social sciences) as being (1) contrastive (so as to be able to compare the fact being questioned with other alternatives or foils); (2) selective (of what information among many to attribute causality of an event to); and (3) social (implying that the explainer must be able to leverage the mental model of the explainee while engaging in the explanation process). In our recent work [7] on explanation generation for planners, we expressed similar sentiments by arguing that the explanation process towards end users "cannot be a soliloquy" but rather a process of "model reconciliation" during which the system tries to bring the mental model (*social property*) of the user on the same page with respect to the plan being explained. We addressed the *contrastive property* by ensuring optimality of the plan being explained in the updated human mental model, and the *selectivity property* by computing the minimum number of updates required to realize the contrast.

The robotics and planning community has indeed seen active research in dealing with such model differences, such as in expressing incapability [22]–[25], communicating misunderstandings about the robot's capabilities [7], [8] or even lying [26] and augmenting new goals [27]. A concise survey of the model reconciliation process can be found in [28].

In related threads [29] of work, we have looked at the notion of "explicable" planning [30], [31] which circumvents the need for explanations by instead having the robot sacrifice optimality in its own model and produce plans that are as close to optimal as possible in the mental model of the human. Of course, such plans may be too costly or even infeasible from the robot's perspective. As such, the process of explicability and explanations form a delicate balancing act [8] during the deliberative process and forms a basis of an augmentative theory [32] of planning for an automated agent.

The process of explanations and explicability for task plans, in general, is also a harder process than in motion planning (c.f. recent works on "legibility" [33] and "verbalization" [34]) where acceptable behavior can be understood in terms of simple rules (e.g. minimizing distance to shortest path). In the case of task planning, human mental models are harder to acquire and thus must be *learned* [30]. Further, given a mental model of the user, it is still a challenge on how to leverage that model in the explanation process, keeping in mind the cognitive abilities and implicit processes and preferences of the human in the loop that are often very hard, or even impossible, to codify precisely in the task model itself. Evaluation of learned mental models is out of scope[1] of the current discussion, though readers are encouraged to refer to [29], [30] for related studies. In this paper, we will focus only on known models, and explore how humans respond to these techniques in situations where these models diverge.

III. Glossary of Terms

Existing teamwork literature [37] on human-human and human-animal teams has identified characteristics of effective teams – in terms of shared mental models [38], [39] that contribute to team situational awareness [40] and interaction [41]. Thus, it has been argued [10] that the ability to leverage these shared mental models, and reasoning over multiple models at a time, during the decision making process is critical to the effective design of cognitive robotic agents for teaming with humans. The multi-model setting is illustrated in Figure 1a in the context of a search and rescue scenario (more on this later in Section IV) where the map of the environment shared across the robot and its operator diverge in course of operations. When making plans in such scenarios, the robot can choose to either (1) conform to human expectations, potentially sacrificing optimality in the process; or (2) preserve optimality and explain its plan (which may thus be inexplicable) in terms of the model differences (that causes

[1] The evaluations, of course, have the same assumptions (and limitations) as the original works on model reconciliation. However, there has been follow up work that relaxes those assumptions (e.g. conformant explanations [35] for model uncertainty and hierarchical explanations [36] for varying degrees of information) without affecting the output of model reconciliation. Thus, our experimental setup remains valid.

this inexplicability). As explained before, the former process is described as explicable planning, while the latter is referred to as explanations as model reconciliation.

A. Explicable Plans

Let the model (which includes beliefs or state information and desires or goals as well as the action model) that the robot is using to plan be given by $\mathcal{M}_R$ and the human's understanding of the same be given by $\mathcal{M}_R^H$. Further, let $\pi^*(\mathcal{M}_R)$ and $\pi^*(\mathcal{M}_R^H)$ be the optimal plans in the respective models, and $C_{\mathcal{M}}(\cdot)$ be the (cost) function denoting the goodness of a plan in a model $\mathcal{M}$. When $\mathcal{M}_R^H \neq \mathcal{M}_R$, it is conceivable that $C_{\mathcal{M}_R^H}(\pi^*(\mathcal{M}_R)) > C_{\mathcal{M}_R^H}(\pi^*(\mathcal{M}_R^H))$ which constitutes an inexplicable behavior from the perspective of the human.

In explicable planning, the robot produces a plan π such that $C_{\mathcal{M}_R^H}(\pi) \approx C_{\mathcal{M}_R^H}(\pi^*(\mathcal{M}_R^H))$, i.e. an explicable plan is equivalent (or as close as possible) to the human expectation.

B. Plan Explanations as Model Reconciliation

Instead, the robot can stay optimal in its own model, and explain the reasons, i.e. model differences, that causes its plan to be suboptimal in the human's mental model.

The Model Reconciliation Problem (MRP) involves the robot providing an explanation or model update $\mathcal{E}$ to the human so that in the new updated human mental model $\widehat{\mathcal{M}}_R^H$ the original plan is optimal (and hence explicable), i.e. $C_{\widehat{\mathcal{M}}_R^H}(\pi^*(\mathcal{M}_R^H)) = C_{\widehat{\mathcal{M}}_R^H}(\pi^*(\mathcal{M}_R))$[2].

Of course, there may be many different types of these explanations, as explained below (terms reused from [7]).

1) Model Patch Explanations (MPE): Providing the entire model difference as a model update is a trivial solution. It satisfies the optimality criterion but may be too large when the robot has to operate with reduced communication bandwidth. It can also cause loss of situational awareness and increased cognitive load of the human by providing too much information that is not relevant to the plan being explained.

2) Plan Patch Explanations (PPE): These restrict model changes to only those actions that appear in the plan. These do not satisfy the optimality criterion but ensure the executability of the given plan instead. Further, they may still contain information that is not relevant to explaining the original robot plan as opposed to the human expectation or foil.

In this paper, we use a specific variant of PPE which contrasts executability with a particular expected human plan. Thus it may still not preserve optimality, but retains the contrastive property of an explanation.

3) Minimally Complete Explanations (MCE): These explanations, on top of satisfying the optimality condition, also enforce $\min \mathcal{E}$. This means MCEs not only make sure that the plan being explained is optimal in the updated model but also it is the minimum set of updates required to make this happen. This is especially useful in reducing irrelevant information during the explanation process both from the perspective of the human as well as the robot when communication is expensive.

4) Minimally Monotonic Explanations (MME): Interestingly, MCEs can become invalid when combined, i.e. when multiple plans are being explained, the current MCE can make a previous one violate the optimality constraint. This leads to the notion of MMEs which guarantee that an explanation is always valid regardless of other plans being explained in the future (while at the same time revealing as little information as possible). This is especially useful in long term interactions and is out of scope of the current study.

C. Balancing Explicability and Explanations

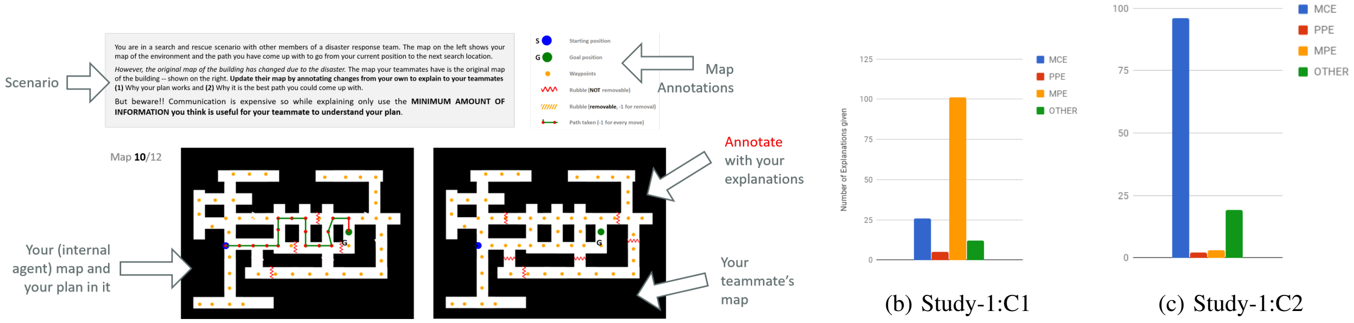
Finally, as mentioned before, these ideas can come together whereby an agent can choose to trade off the cost of explanations versus the cost of producing explicable plans by performing model space search during the plan generation process [8]. In the following studies, we simulate such an agent that generates plans that are either optimal in its own model (with an associated MCE, MPE or PPE) or explicable or somewhere in between (with an associated MCE).

IV. Testbed: The USAR Domain

An application where such multi-model formulations are quite useful is in typical [6] Urban Search And Reconnaissance (USAR) tasks where a remote robot is put into disaster response operation often controlled partly or fully by an external human commander who orchestrates the entire operation. The robot's job in such scenarios is to infiltrate areas that may be otherwise harmful to humans, and report on its surroundings as and when required / instructed by the external supervisor. The external usually has a map of the environment, but this map may no longer be accurate in the event of the disaster – e.g. new paths may have opened up, or older paths may no longer be available, due to rubble from collapsed structures like walls and doors. The robot (internal) however may not need to inform the external of all these changes so as not to cause information overload of the commander who may be otherwise engaged in orchestrating the entire operation. The robot is thus delegated high level tasks but is often left to compute the plans itself since it may have a better understanding of the environment. However, the robot's actions also contribute to the overall situational awareness of the external, who may require explanations on the robots plans when necessary. As such, simulated USAR scenarios provide an ideal testbed for developing and evaluating algorithms for effective human-robot interaction. Figure 1b illustrates our setup (c.f. https://youtu.be/40Xol2GY7zE) In the current study, we only simulate the interface to the external (Section VI).

Differences in the models of the human and the robot can manifest in many forms (e.g. the robot may have lost some capability or its goals may have changed). In our setup, we deal with differences in the map of the environment as available to the two agents – these can be compiled to differences only in the initial state of the planning problem (the human

[2] We refer to this constraint as "the optimality condition" and the explanations that satisfy this condition are called complete explanations.



You are in a search and rescue scenario with other members of a disaster response team. The map on the left shows your map of the environment and the path you have come up with to go from your current position to the next search location.

However, the original map of the building has changed due to the disaster. The map your teammates have is the original map of the building -- shown on the right. **Update their map by annotating changes from your own to explain to your teammates (1)** Why your plan works and **(2)** Why it is the best path you could come up with.

But beware!! Communication is expensive so while explaining only use the **MINIMUM AMOUNT OF INFORMATION you think is useful for your teammate to understand your plan**.

(a) Interface for Study-1

(b) Study-1:C1

(c) Study-1:C2

(d) Explanation counts for Study-1:C:1-2.

Fig. 2: In Study-1, participants assumed the role of the internal agent and were asked to explain their plan to a teammate with a possibly different model or map of the world.

model has the original unaffected model of the world). This makes no difference to the underlying explanation generation algorithm [7] which treats all model changes equally.

While the availability of models (as required by all the algorithms in III) may be a strong assumption in some cases, in domains such as USAR this is, indeed, the case since teams in such scenarios start off with a shared model (e.g. blueprint of a building). The USAR domain is also ideal for visualizing to non-expert participants, in comparison to, for example, logistics-type domains which should ideally be evaluated by experts. This became an important factor while designing the user studies. The USAR domain is thus at once close to motion planning as easily interpreted by non-experts but also incorporates typical combinatorial aspects of task plans such as preconditions and effects in terms of rubble removal, collapsed halls, etc. and relevant abilities of the robot.

V. Study – 1

The first part of the study aims to develop an understanding of how humans respond to the task of generating explanations, i.e. if left to themselves, humans preferred to generate explanations similar to the ones enumerated in Section III-B. To test this, we asked participants to assume the role of the internal agent in the explanation process and explain their plans with respect to the faulty map of their teammate. Specifically, we set out to test the following hypothesis –

H1. When asked to, participants would leverage model differences as a key ingredient for explanations.

  H1a. Explanation generated by participants would demonstrate contrastiveness. Thus, PPE type explanations would be overlooked in favor of complete solutions (MCEs and MPEs) when there are multiple competing hypothesis for the human.

H2. Participants would like to minimize the content of the explanation by removing details that are not relevant to the plan being explained.

  H2a. Explanations generated by humans would be closer to MCEs than MPEs.

  H2b. This should be even more significant if restrictions are placed on communication.

As a result of this study, we intend to identify to what extent explanation types for task planning studied in existing literature (Section III-B) that claim to build upon principles of explanations in human-human interactions studied in social sciences [20] truly reflect human intuition.

Note that we primed the subjects to annotate changes in the map, while giving them the opportunity to –

1. Provide more than annotations (and we did find other interesting kinds of explanations emerge as we discuss later in Section VII)
2. Comment on the sufficiency and necessity of such explanations (as we report in Section V-B)

The reason for this choice was because in the work being evaluated here (c.f. Section III), communicating model differences has been considered to be the starting point of the explanation process. So we start from that assumption and evaluate whether the kinds of explanations introduced in existing literature – MCE / MPE / PPE / etc. – are actually useful. Additionally, this setup also helps to re-contextualize the real importance of model difference in the explanation process in light of reasons explained in (1) and (2) above.

A. Experimental Setup

Figure 2a shows an example map and plan provided to a participant. On left side, the participant is shown the actual map along with the plan, starting position and the goal. The panel on the right shows the map that is available to the explainee. The maps have rubbles (both removable and non-removable) blocking access to certain paths. The maps may disagree as to the locations of the debris. The participants were told that they need to convince the explainee of the correctness and optimality of the given plan by updating the latter's maps with annotations they felt were relevant in achieving that goal. We ran the study with two conditions –

C1. Here the participants were asked to ensure, via their explanations, that their plan was (1) correct and (2) optimal in the updated model of their teammate; and

C2. Here, in addition to C1, they were also asked to use the minimal amount of information they felt was needed to achieve the condition in C1.

Each participant was shown how to annotate (not an explanation) a sample map and was then asked to explain 12 different plans using similar annotations. After each participant was finished with their assignment, they were asked the following subjective questions –

Q1. Providing map updates were necessary to explain my plans.
Q2. Providing map updates were sufficient to explain my plans.
Q3. I found that my plans were easy to explain.

The answers to these questions were measured using a five-point Likert scale. The answers to the first two questions will help to establish whether humans considered map updates (or in general updates on the model differences) at all necessary and/or sufficient to explain a given plan. The final question measures whether the participants found the explanation process using model differences tractable. It is important to note that in this setting we do not measure the efficacy of these explanations (this is the subject of Study-2 in Section VI). Rather we are trying to find whether a human explainer would have naturally participated in the model reconciliation approach during the explanation process.

In total, we had 12 participants for condition C1 and 10 participants for condition C2 including 7 female and 18 male participants between the age range of 18-29 (data corresponding to 5 participants who misinterpreted the instructions had to be removed, 2 participants did not reveal their demographics). Participants for the study were recruited by requesting the department secretary to send an email to the student body to ensure that they had no prior knowledge about the study or its relevance. Each participant was paid $10 for taking part.

B. Results

The results of the study are presented in Figures 2d, 3 and 4. We summarize some of the major findings below –

Figure 2d – The first hypothesis we tested was whether the explanations generated by the humans matched any of the explanation types discussed in Section III-B. We did this by going through all the individual explanations provided by the participants and then categorizing each explanation to one of the four types, namely MCE, PPE, MPE or Other (the ”other” group contains explanations that do not correspond to any of the predefined explanation types – more on this later in Section VII). Figure 2b shows the number of explanations of each type that were provided by the participants of C1. The graph shows a clear preference for MPE, i.e. providing all model differences. A possible reason for this may be since the size of MPEs for the given maps were not too large (and participants did not have time constraints). Interestingly, in C2 we see a clear shift in preferences (Figure 2c) where most participants ended up generating MCE style explanations. This means at least for scenarios where there are constraints on communication, the humans would prefer generating MCEs as opposed to explaining all the model differences.

These findings are consistent with H1, with very few of the explanations in type “Other” (Figure 2d). This is also backed up by answers to subjective questions Q1 and Q2 above. Further, the preference of MPE/MCE over PPE (H1a) is quite stark. Contrary to H2a, participants seemed to have preferred full model explanation (MPE) in C1 condition which is surprising. However, results of C2 condition are more aligned with H2b, even though we expected to see similar trend (if not as strong) in C1 condition as well.

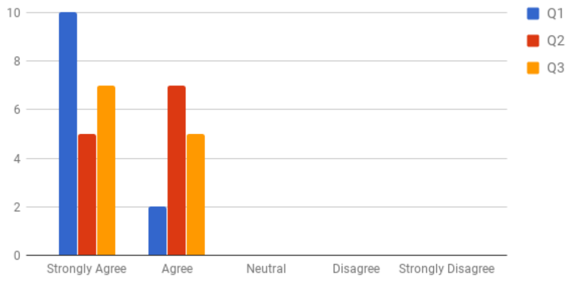


Fig. 3: Subjective responses of participants in Study-1:C1.

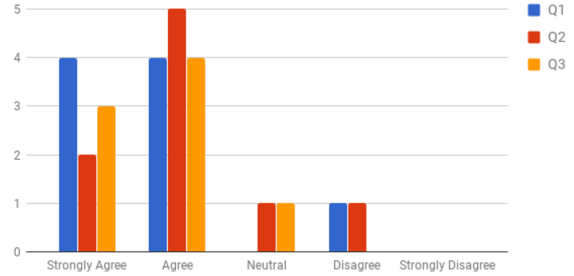


Fig. 4: Subjective responses of participants in Study-1:C2.

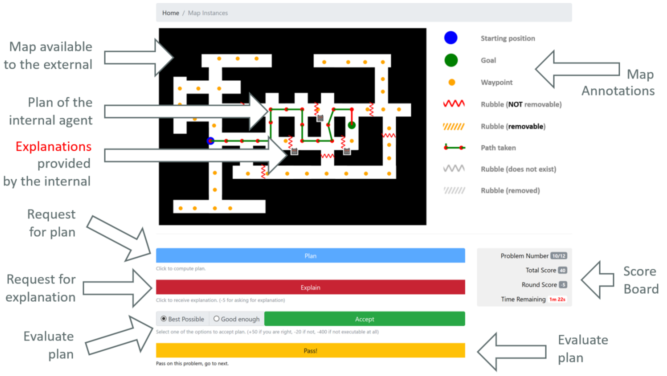
Figures 3 and 4 – These show the results of the subjective questions for C1 and C2 respectively. Interestingly, in C1, while most people agreed on the necessity of explanations in the form of model differences, they were less confident regarding the sufficiency of such explanations. In fact, we found that many participants left additional explanations in their worksheet in the form of free text (we discuss some of these findings in Section VII). In C2, we still see that more people are convinced about the necessity of these explanations than sufficiency. But we see a reduction in the confidence of the participants, which may have been caused by the additional minimization constraints.

VI. Study – 2

Here we study how different kinds of explanations outlined in Section III-B are perceived or evaluated when they are presented to the participants. This study was designed to provide clues to how humans comprehend explanations when provided to them in the form of model differences. Specifically, we intend to evaluate the following hypothesis, in line with the intended properties of each of the explanation and plan types in existing literature (Section III-B) –

H1. Participants would be able to identify optimality of a plan given an MPE or an MCE.

H2. Participants would be able to identify executability but possible suboptimality of a plan given a PPE.



Click to compute plan.

Click to receive explanation. (-5 for asking for explanation)

Select one of the options to accept plan. (+50 if you are right, -20 if not, -400 if not executable at all)

Pass on this problem, go to next.

Fig. 5: Interface for Study-2 where participants assumed the role of the external commander and evaluated plans provided by the internal robot. They could request for plans and explanations to those plans (if not satisfied) and rate them as optimal or suboptimal or (if unsatisfied) can chose to pass.

H3. Participants would not ask for explanations when presented with explicable plans.

As a result of this study, we intend to validate whether desired properties of explanations for task planning designed by following norms and principles outlines in the social sciences in the context of human-human interactions [20] do actually carry over for human-robot interactions.

A. *Experimental Setup*

During this study, participants were incentivized to make sure that the explanation does indeed help them understand the optimality and correctness of the plans in question by formulating the interaction in the form of a game.

Figure 5 shows a screenshot of the interface. The game displays to each participant an initial map (which they are told may differ from the robot's actual map), the starting point and the goal. Once the player asks for a plan, the robot responds with a plan illustrated as a series of paths through waypoints highlighted on the map. The goal of the participant is to identify if the plan shown is optimal or just satisficing. If the player is unsure of the path, they can ask for an explanation from the robot. The explanation is provided to the participant in the form of a set of model changes in the player's map. If the player is still unsure, they can click on the pass button to move to the next map.

The scoring scheme for the game is as follows. Each player is awarded 50 points for correctly identifying the plan as either optimal or satisficing. Incorrectly identifying an optimal plan as suboptimal or vice versa would cost them 20 points. Every request for explanation would further cost them 5 points, while skipping a map does not result in any penalty. The participants were additionally told that selecting an inexecutable plan as either feasible or optimal would result in a penalty of 400 points. Even though there were no actual incorrect plans in the dataset, this information was provided to deter participants from taking chances with plans they did not understand well.

Each participant was paid $10 dollars and received additional bonuses based on the following payment scheme –

- Scores higher than or equal to 540 were paid $10.
- Scores higher than 540 and 440 were paid $7.
- Scores higher than 440 and 340 were paid $5.
- Scores higher than 340 and 240 were paid $3.
- Scores below 240 received no bonuses.

The scoring systems for the game was designed to ensure

- Participants should only ask for an explanation when they are unsure about the quality of the plan (due to small negative points on explanations).
- Participants are incentivized to identify the feasibility and optimality of the given plan correctly (large reward and penalty on doing this wrongly).

Each participant was shown a total of 12 maps (same maps as in Study–1). For 6 of the 12 maps, the player was assigned the optimal robot plan, and when they asked for an explanation, they were randomly shown either MCE, PPE or MPE explanation with regards to the robot model (as per algorithms described in Section III-B). For the rest of the maps, participants could potentially be assigned a plan that is optimal in the human model (i.e. an explicable plan) or somewhere in between as introduced in Section III-C (referred to henceforth as the balanced plan) in place of the robot optimal plan[3]. The participants that were assigned the optimal robot plan were still provided an MCE, PPE or MPE explanation, otherwise they were provided either the shorter explanation (for balanced plans) or an empty explanation (for the explicable plan). Also note that for 4 out of the 12 maps the PPE explanation cannot prove the optimality of the plan.

At the end of the study, each participant was presented with a series of subjective questions. The responses to each question were measured on a five-point Likert scale.

```
Q1. The explanations provided by the robot was helpful.
Q2. The explanations provided by the robot was easy to understand.
Q3. I was satisfied with the explanations.
Q4. I trust the robot to work on its own.
Q5. My trust in the robot increased during the study.
```

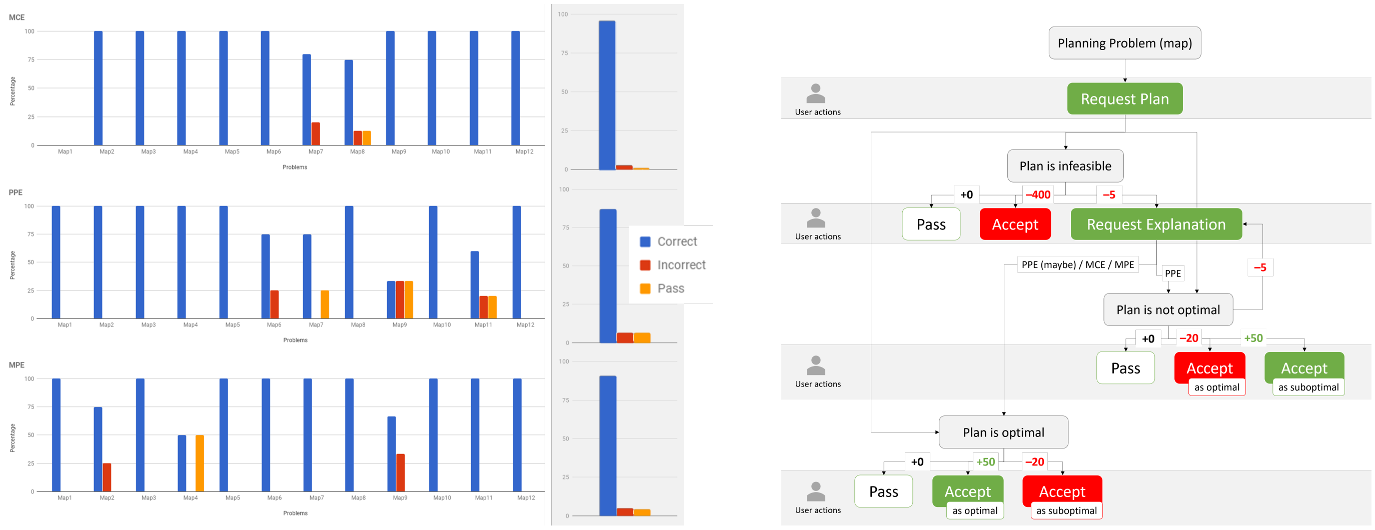
In total, we had 27 participants for Study–2, including 4 female and 22 male participants between the age range of 19-31 (1 participant did not reveal their demographic).

B. *Results*

The results of the study are presented in Figures 6, 7 and 8. We summarize some of the major findings below –

Figure 6 – As we mentioned before, the goal of this study is to identify if explanations in the form of model reconciliation can convey to humans the optimality and correctness of plans. Here, each participant was shown the 12 maps from Study-1 and each map was assigned a random explanation type (and in some cases different plans). We wanted to identify whether the participants that asked for explanations were able to come up with the correct conclusions. This means that the subjects who

[3] Note that of the 6 maps, only 3 had both balanced as well as explicable plans, the rest either had a balanced plan or the optimal human plan.



(a)

(b)

Fig. 6: Percentage of times (inset 6a) different explanations (i.e. MCE / MPE / PPE) led to correct decision on the human's part in each problem (the aggregated result is shown on the right). A "correct decision" involves recognizing optimality of the robot plan on being presented an MCE or MPE, and optimality or executability (as the case may be) in case of a PPE. Inset 6b illustrates this flow of logic in the experimental setup.

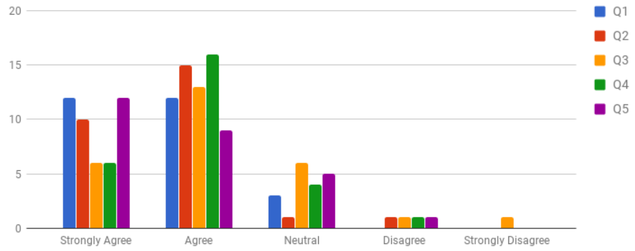


Fig. 7: Subjective responses of participants in Study–2.

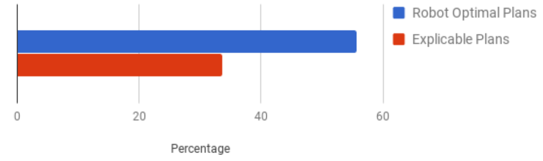


Fig. 8: Percentage of times explanations were sought for in Study–2 when participants presented with explicable plans versus robot optimal plans with explanations.

asked for MCE and MPE were able to correctly identify the plans as optimal, while the people who received PPE were able to correctly classify the plan to either optimal or satisficing (i.e. for all but 5 maps PPE is enough to prove optimality).

Figure 6a shows the statistics of the selections made by participants who had requested an explanation. The right side of the graph shows the percentage (for every map instance) of participants who selected the correct options (marked in blue), the incorrect ones (marked in red) or simply passed (marked in orange), while the left side shows the average across all 12 maps. We notice that in general people were overwhelmingly able to identify the correct choice. Even in the case of PPEs, where the explanations only ensured correctness (map instances 1, 2, 3, 8 and 11) the participants were able to make the right choice. This is consistent with H1 and H2 and demonstrates that explanations in the form of model reconciliation are a viable means of conveying the correctness and optimality of plans – i.e. participants can differentiate between complete and non-complete explanations.

Figure 7 – These conclusions are further supported by results from the subjective questionnaire (Figure 7). Most people seem to agree that the explanations were helpful and easy to understand. In fact, the majority of people strongly agreed that their trust of the robot increased during the study.

Figure 8 – We were also curious (H3) about the usefulness of explicable plans (that are optimal in human's model), i.e. if the subjects still asked for explanations when presented with explicable plans. Figure 8 shows the percentage of times subjects asked for explanations when presented with explicable versus robot optimal plans. The rate of explanations is considerably less in case of explicable plans as desired. This matches the intuition behind the notion of plan explicability as a viable means (in addition to explanations) of dealing with model divergence in human-in-the-loop operation of robots.

It is interesting to see that in Figure 8 about a third of the time participants still asked for explanations even when the plan was explicable, and thus optimal in their map. This is an artifact of the risk-averse behavior incentivized by the gamification of the explanation process. This is to make sure that participants were sufficiently invested in the outcome as well as mimic the high-stakes nature of USAR settings to accurately evaluate the explanations. It is also an indication

TABLE I: Summary of results.

Study	Hypothesis	Outcome	Comments
Study-1	H1	✓	Participants largely agreed that model reconciliation was a necessary and sufficient part of the explanation process.
	H1a	✓	Participants preferred explanations that are complete, and preserve contrastive property across multiple hypothesis.
	H2	✗	Participants did not care to minimize size of explanations, i.e. exclude irrelevant details.
	H2a	✗	Explanations in the free form condition were largely of MPE type.
	H2b	✓	Participants did generate MCEs when their communication capability was explicitly restricted.
Study-2	H1	✓	Participants could identify plan optimality in response to complete explanations.
	H2	✓	Participants could identify suboptimality for incomplete explanations.
	H3	✓ / ?	Some participants asked for explanations even for explicable plans, though the majority did not.

of the cognitive burden on the humans who may not be (cost) optimal planners. While this is consistent with the spirit of H3, the finding is also somewhat indicative of the limitations of plan explicability as it is defined in existing literature at the moment [8]. Thus, going forward, the objective function should incorporate the cost or difficulty of analyzing the plans and explanations from the point of view of the human in addition to the current costs of explicability and explanations modeled from the perspective of the robot.

Interestingly, the participants also did not ask for explanations around 40% of the time (c.f. Figure 8) when they "should have" (i.e. suboptimal plan in the human model) according to the theory of model reconciliation. We noticed no clear trend here (e.g. decreasing rate for explanations asked due to increasing trust). This was most likely due to limitations of inferential capability of humans and a limitation of the existing formulation of model reconciliation as well.

Also note that balanced plans are indistinguishable from the point of view of the human and are more useful to the robot for trading of explanation and explicability costs. Hence, we did not expand on further results on balanced plans so as not to distract from the main focus of the paper which is to evaluate explanations as model reconciliation. A more detailed account of Figure 8 can be found in [8].

VII. Discussions

As mentioned before, there were instances where participants in Study 1 generated explanations that are outside the scope of any of the explanations discussed in Section III-B. These are marked as "Other" in Figure 2d. In the following we note three of these cases that we found interesting –

a) Post-hoc explanations: Notice that parts of an MCE that actually contribute to the executability of a given plan may not be explained in post-hoc situations where the robot is explaining plans that have *already been done* as opposed to plans that are being proposed for execution. The rationale behind this is that if the human sees an action, that would not have succeeded in his model, actually end up succeeding (e.g. the robot had managed to go through a corridor that was blocked by rubble) then he can rationalize that event by updating his own model (e.g. there must not have been a rubble there). This seems to be a viable approach to further reduce size (c.f. selective property of explanations [20]) of explanations in a post-hoc setting, and is out of scope of explanations developed in [7].

b) Identification of Explicit Foils: Identification of explicit foils [7] can help reduce the size of explanations. In the explanations introduced in Section III-B the foil was implicit – i.e. *why this plan as opposed to all other plans*. However, when the implicit foil can be estimated (e.g. top-K plans expected by the human) then explanations may only include information on why the plan in question is better than the other options (which are either costlier or not executable). Some participants provided explanations contrasting such foils in terms of (and in addition to) the model differences.

c) Cost-based reasoning: Finally, a kind of explanation that was attempted by some participants involved a cost analysis of the current plan with respect to foils (in addition to model differences, as mentioned above). Such explanations have been studied extensively in previous planning literature [16], [19] and seems to be still relevant for plan explanations on top of the model reconciliation process.

VIII. Conclusion

The paper details the results of studies aimed to evaluate the effectiveness of plan explanations in the form of model reconciliation. Through this study, we aimed to validate whether explanations in the form of model reconciliation (in its various forms) suffice to explain the optimality and correctness of plans to the human in the loop. We also studied cases where participants were asked to generate explanations in the form of model changes, to see if explanations generated by the humans align with any of the explanations identified in existing literature. The results of the study (summarized in Table I) seem to suggest that humans do indeed understand explanations of this form and believe that such explanations are necessary to explain plans. In future work, we would like to investigate how explanations can be adapted for scenarios where the robot is expected to interact with the humans over a long period of time and how such interactions affect the dynamics of trust and teamwork. Recent work on explanations using *model of self* [42] can also provide interesting ways of *abstracting* [43] model information.

References

[1] David Gunning, “Explainable Artificial Intelligence (XAI),” https://goo.gl/geab6t, 2016, DARPA Program.

[2] David W. Aha, “Workshop on Explainable Artificial Intelligence (XAI),” https://goo.gl/B2iby2, 2017-18, IJCAI.

[3] John Frank Weaver, “Artificial Intelligence Owes You an Explanation,” https://goo.gl/a55LK9, 2017, Slate.

[4] P. Langley, B. Meadows, M. Sridharan, and D. Choi, “Explainable Agency for Intelligent Autonomous Systems,” in *AAAI/IAAI*, 2017.

[5] B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a “right to explanation”,” *ArXiv*, 2016.

[6] C. E. Bartlett, “Communication between Teammates in Urban Search and Rescue,” *Thesis*, 2015.

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

[8] T. Chakraborti, S. Sreedharan, and S. Kambhampati, “Balancing Explicability and Explanation in Human-Aware Planning,” in *AAMAS*, 2018.

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

[10] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang, “AI Challenges in Human-Robot Cognitive Teaming,” *CoRR*, vol. abs/1707.04775, 2017.

[11] T. Lombrozo, “The structure and function of explanations,” *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464 – 470, 2006.

[12] ——, “Explanation and abductive inference,” *Oxford handbook of thinking and reasoning*, pp. 260–276, 2012.

[13] J. D. Moore and W. R. Swartout, “Explanation in Expert Systems: A survey,” University of Southern California Maria Del Rey Information Sciences Institute, Tech. Rep., 1988.

[14] S. Kambhampati, “A classification of plan modification strategies based on coverage and information requirements,” in *AAAI Spring Symposium on Case Based Reasoning*, 1990.

[15] F. Sørmo, J. Cassens, and A. Aamodt, “Explanation in case-based reasoning–perspectives and goals,” *Art. Int. Review*, 2005.

[16] D. E. Smith, “Planning as an Iterative Process,” *AAAI*, 2012.

[17] R. Howey, D. Long, and M. Fox, “Val: Automatic plan validation, continuous effects and mixed initiative planning,” in *ICTAI*, 2004.

[18] S. Sohrabi, J. A. Baier, and S. A. McIlraith, “Preferred explanations: Theory and generation via planning,” in *AAAI*, 2011.

[19] M. Fox, D. Long, and D. Magazzeni, “Explainable Planning,” in *IJCAI XAI Workshop*, 2017.

[20] T. Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences,” *AIJ*, 2018.

[21] T. Miller, P. Howe, and L. Sonenberg, “Explainable AI: Beware of Inmates Running the Asylum,” in *IJCAI XAI Workshop*, 2017.

[22] V. Raman, C. Lignos, C. Finucane, K. C. Lee, M. Marcus, and H. Kress-Gazit, “Sorry dave, i’m afraid i can’t do that: Explaining unachievable robot tasks using natural language,” in *Robotics: Science and Systems (RSS)*, 2013.

[23] V. Raman and H. Kress-Gazit, “Towards minimal explanations of unsynthesizability for high-level robot behaviors,” in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 757–762.

[24] G. Briggs and M. Scheutz, ““Sorry, I can’t do that”: Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions,” in *2015 AAAI Fall Symposium Series*, 2015.

[25] M. Kwon, S. Huang, and A. Dragan, “Expressing Robot Incapability,” in *HRI*, 2018.

[26] T. Chakraborti and S. Kambhampati, “Algorithms for the Greater Good! On Mental Modeling and Acceptable Symbiosis in Human-AI Collaboration,” *AIES/AAAI*, 2019.

[27] Z. Chen and Y. Zhang, “Explain by Goal Augmentation: Explanation Generation as Inverse Planning,” in *RSS Workshop on on Adversarial Robotics*, 2018.

[28] T. Chakraborti, S. Sreedharan, and S. Kambhampati, “Human-Aware Planning Revisited: A Tale of Three Models,” in *IJCAI-ECAI XAI / ICAPS XAIP Workshops*, 2018.

[29] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati, “Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior,” *CoRR*, vol. abs/1811.09722, 2018. [Online]. Available: http://arxiv.org/abs/1811.09722

[30] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, “Plan Explicability and Predictability for Robot Task Planning,” in *ICRA*, 2017.

[31] A. Kulkarni, T. Chakraborti, Y. Zha, S. G. Vadlamudi, Y. Zhang, and S. Kambhampati, “Explicable Robot Planning as Minimizing Distance from Expected Behavior,” *CoRR*, vol. abs/1611.05497, 2016.

[32] H. Mercier and D. Sperber, “Why Do Humans Reason? Arguments for an Argumentative Theory,” *Behavioral and Brain Sciences*, 2010.

[33] A. Dragan, K. Lee, and S. Srinivasa, “Legibility and predictability of robot motion,” in *HRI*, 2013.

[34] V. Perera, S. P. Selveraj, S. Rosenthal, and M. Veloso, “Dynamic generation and refinement of robot verbalization,” in *RO-MAN*, 2016.

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

[36] S. Sreedharan, S. Srivastava, and S. Kambhampati, “Hierarchical Expertise-Level Modeling for User Specific Robot-Behavior Explanations,” in *IJCAI*, 2018.

[37] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, “Interactive team cognition,” *Cognitive science*, vol. 37, no. 2, pp. 255–285, 2013.

[38] J. Cannon-Bowers, E. Salas, and S. Converse, “Shared mental models in expert team decision making,” *Current issues in individual and group decision making*.

[39] J. E. Mathieu, T. S. Heffner, G. F. Goodwin, E. Salas, and J. A. Cannon-Bowers, “The influence of shared mental models on team process and performance,” *Journal of Applied Psychology*, 2000.

[40] J. Gorman, N. Cooke, and J. Winner, “Measuring team situation awareness in decentralized command and control environments.” *Ergonomics*, vol. 49, pp. 1312–1325, 2006.

[41] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. Duran, “Interactive team cognition,” *Cognitive Science*, 2013.

[42] B. Hayes and J. A. Shah, “Improving robot controller transparency through autonomous policy explanation,” in *HRI*, 2017.

[43] S. Sreedharan, S. Srivastava, and S. Kambhampati, “Hierarchical Expertise-Level Modeling for User Specific Robot-Behavior Explanations,” *IJCAI*, 2018.