Balancing Explicability and Explanations for Human-Aware Planning

Tathagata Chakraborti1‡, Sarath Sreedharan2∗ and Subbarao Kambhampati2

1IBM Research AI, Cambridge MA 02142 USA
2Arizona State University, Tempe AZ 85281 USA
tchakra2@ibm.com, {ssreedh3, rao}@asu.edu

Abstract

Human-aware planning involves generating plans that are explicable as well as providing explanations when such plans cannot be found. In this paper, we bring these two concepts together and show how an agent can achieve a trade-off between these two competing characteristics of a plan. In order to achieve this, we conceive a first of its kind planner MEGA that can augment the possibility of explaining a plan in the plan generation process itself.

We situate our discussion in the context of recent work on explicable planning and explanation generation, and illustrate these concepts in two well-known planning domains, as well as in a demonstration of a robot in a typical search and reconnaissance task. Human factor studies in the latter highlight the usefulness of the proposed approach.

1 Introduction

It is often useful for an agent while interacting with a human to use, in the process of its deliberation, not only its own model $M_R$ of the task, but also the model $M_R^h$ that the human thinks it has (as shown in Figure 1). This mental model [Chakraborti et al., 2017a] is in addition to the task model of the human $M_H^r$ (denoting their beliefs, intentions and capabilities). This is, in essence, the fundamental thesis of the recent works on plan explanations as model reconciliation [Chakraborti et al., 2017b] and explicable planning [Zhang et al., 2017] and is in addition to the originally studied human-aware planning (HAP) problems where actions of the human (i.e. the human task model and a robot’s belief of it) are involved in the planning process. The need for explicability and explanations occur when these two models – $M_R$ and $M_R^h$ – diverge. This means that the optimal plans in the respective models – $\pi_{M_R}^*$ and $\pi_{M_R^h}^*$ – may not be the same and hence optimal behavior of the robot in its own model may seem inexplicable to the human.

– In explicable planning, the robot produces a plan $\bar{\pi}$ that is closer to the human’s expected plan, i.e. $\bar{\pi} \approx \pi_{M_R^h}^*$.

– During plan explanation (as model reconciliation), it updates the mental model to an intermediate model $M_R^\bar{\pi}$ in which the robot’s original plan is equivalent (with respect to a metric such as cost or similarity) to the optimal one and hence explicable, i.e. $\pi_{M_R^\bar{\pi}}^* \equiv \pi_{M_R}^*$.

Until now, these two processes of plan explanations and explicability have remained separate in so far as their role in an agent’s deliberative process is considered - i.e. a planner either generates an explicable plan to the best of its ability or it produces explanations of its plans where they required. However, there are situations where a combination of both provide a much better course of action – if the expected human plan is too costly in the planner’s model (e.g. the human might not be aware of some safety constraints) or the cost of communication overhead for explanations is too high (e.g. limited communication bandwidth). Consider, for example, a human working with a robot that has just received a software update

Figure 1: The explicability versus explanation trade-off in human-aware planning. The planner accounts for the human’s model of itself in addition to its own model – it can either bring the former closer to its own using explanations via the model reconciliation process so that an otherwise inexplicable plan now makes sense in the human’s updated model and/or it can produce explicable plans which are closer to the human’s expectation of optimality.

This is an extended version of an abstract that appeared previously at AAMAS 2018 [Chakraborti et al., 2018].

†Authors marked with ‡ contributed equal.
allowing it to perform new complex maneuvers. Instead of directly trying to conceive all sorts of new interactions right away that might end up spooking the user, the robot could instead reveal only certain parts of the new model while still using its older model (even though suboptimal) for the rest of the interactions so as to slowly reconcile the drifted model of the user. This is the focus of the current paper where we try to attain the sweet spot between plan explanations and explicability during the planning process. To this end:

1. We develop a first of its kind planner that can envisage possible explanations required of its plans and incorporate these considerations in the planning process itself.

   1a. We show how this allows us to generate explanations even shorter than the previously proposed “shortest possible” explanations in [Chakraborti et al., 2017b] given a plan.

   1b. Since the explicability problem has been studied in plan space while explanation generation works in model space, a viable solution to the balancing act cannot be a simple combination of the two. Our planner not only computes a plan given a model but also what model to plan in given the human mental model. This also means that, in contrast to explicability-only approaches, we can deal with situations where an explicable plan does not exist by being able to reconcile model differences in the same planning framework.

2. We illustrate the salient features of the algorithm in two well-known planning domains and in human factors studies in a mock search and rescue domain. The empirical evaluations demonstrate the effectiveness of the approach from the robot’s perspective, while the study highlight its usefulness in being able to conform to expected normative behavior.

1.1 Illustrative Example

We illustrate our approach on a robot performing an Urban Search And Reconnaissance (USAR) task – here a remote robot is put into disaster response operation controlled partly or fully by an external human commander. This is a typical USAR setup [Bartlett, 2015] where the robot’s job is to infiltrate areas that may be otherwise harmful to humans, and report on its surroundings as and when required or instructed by the external. The external usually has a map of the environment, but this map is no longer accurate in a disaster setting – 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 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. This requires the robot to reason about the model differences due to changes in the map, i.e. the initial state of the planning problem.

Figure 2 shows a relevant section of the map of the environment where this whole scenario plays out. A video demonstration can be viewed at https://youtu.be/Yzp4FU6Vn0M. The dark marks indicate rubble that has blocked a passage. A lot of rubble cannot be removed. The robot (Fetch), currently located at the position marked with a blue O, is tasked with taking a picture at location marked with an orange O. The commander expects the robot to take the path shown in red, which is no longer possible. The robot has two choices – it can either follow the green path and explain the revealed passageway due to the collapse, or compromise on its optimal path, clear the rubble and proceed along the blue path. The first part of the video demonstrates the plan that requires the least amount of explanation, i.e. the most explicable plan. The robot only needs to explain a single initial state change to make its plan optimal in the updated map of the commander: remove-has-initial-state-clear_path p1 p8

This is an instance where the plan closest to the human expectation, i.e. the most explicable plan, still requires an explanation, which previous approaches in the literature cannot provide. Moreover, in order to follow this plan, the robot must perform the costly clear_passage p2 p3 action to traverse the corridor between p2 and p3, which it could have
avoided in its optimal (green) path. Indeed, the robot’s optimal plan requires the following explanation:

\[
\text{add-has-initial-state-clear\_path p6 p7} \\
\text{add-has-initial-state-clear\_path p7 p5} \\
\text{remove-has-initial-state-clear\_path p1 p8}
\]

By providing this explanation, the robot is able to convey to the human the optimality of the current plan as well as the infeasibility of the human’s expected plan (shown in red).

### 1.2 Related Work

The need for human-aware agents to be able to explain their behavior has received increased attention in recent times [Langley et al., 2017; Rosenthal et al., 2016]. This is highlighted in the success of recent workshops on explainable AI [XAI, 2018] and planning in particular [XAIP, 2018].

Efforts to make planning more “human-aware” have largely focused on incorporating an agent's understanding of the human model into its decision making. Since then the importance of considering the human’s understanding of the agent’s actual model $\mathcal{M}^h$ in the planning process has also been acknowledged, sometimes implicitly [Alami et al., 2014] and later explicitly [Zhang et al., 2017; Chakraborti et al., 2017b]. These considerations allow a human-aware agent to conceive novel and interesting behaviors by reasoning both in the space of plans as well as models.

In the model space, human mental models $\mathcal{M}_h^m$ may be used for explanations [Chakraborti et al., 2017b] while reasoning over the actual task model $\mathcal{M}^H$ can reveal interesting behavior by affecting the state of the human, such as in [Chakraborti et al., 2015]. In the plan space, a human-aware agent can use $\mathcal{M}^H$ and $\mathcal{M}_h^m$ to compute joint plans for teamwork [Talamadupula et al., 2014] or generate behavior that conforms to the human’s preferences [Alami et al., 2006; Alami et al., 2014; Cirillo et al., 2010; Koeckemann et al., 2014; Tomić et al., 2014; Chakraborti et al., 2016] and expectations [Dragan et al., 2013; Zhang et al., 2017; Kulkarni et al., 2019] and create plans that help the human understand the robot’s objectives [Sadigh et al., 2016].

In general, preference modeling looks at constraints on plan generation if the robot wants to contribute to the human utility, while explicability addresses how the robot can adapt its behavior to human expectation (as required by the human mental model). For a detailed treatise of these distinctions, we refer the reader to [Chakraborti et al., 2019a].

### 2 Human-Aware Planning

#### 2.1 Explicable Plans

An explicable solution to an HAP is a plan $\pi$ (1) executable in the robot’s model and (2) closest to the expected (optimal) plan in the human’s model –

\[
\begin{align*}
(1) & \quad \delta_{\mathcal{M}^h}(\mathcal{I}, \pi) \models \mathcal{G}; \\
(2) & \quad C(\pi, \mathcal{M}^h) \approx C^*_{\mathcal{M}^h^r}.
\end{align*}
\]

“Closeness” or distance to the expected plan is modeled here in terms of cost optimality, but in general this can be any metric such as plan similarity. In existing literature [Zhang et al., 2017; Kulkarni et al., 2019] this has been achieved by modifying the search process so that the heuristic that guides the search is driven by the robot’s knowledge of the human mental model. Such a heuristic can be either derived directly [Kulkarni et al., 2019] from the mental model or learned [Zhang et al., 2017] through interactions in the form of affinity functions between plans and their purported goals.

#### 2.2 Plan Explanations

The other approach would be to (1) compute optimal plans in the planner’s model as usual, but also provide an explanation (2) in the form of a model update to the human so that (3) the same plan is now also optimal in the updated mental model. Thus, a solution involves a plan $\pi$ and an explanation $\mathcal{E}$ –

\[
\begin{align*}
(1) & \quad C(\pi, \mathcal{M}^h) = C^*_{\mathcal{M}^h^r}; \\
(2) & \quad \mathcal{M}^l \leftarrow \mathcal{M}^h + \mathcal{E}; \\
(3) & \quad C(\pi, \mathcal{M}^l) = C^*_{\mathcal{M}^h^r}.
\end{align*}
\]

A model update, as indicated by the $+$ operator, may include a correction to the belief (goals or state information) as well as information pertaining to the action model itself, as illustrated in [Chakraborti et al., 2017b]. As a result of this explanation, the human and the agent both agree that the given plan is the best possible the latter could have come up with. Note that whether there is no solution in the human model, or just a different one, does not make any difference. The solution is still an explanation so that the given plan is the best possible in the updated human model. On the other hand, if there is no plan in the robot model, the explanation ensures that there is no plan in the updated human model either.

\footnote{Note that this does not assume that humans use an explicitly represented symbolic domain to plan. The robot only uses this to represent the information content of that model. It cannot, of course, have direct access to it. There is extensive work on learning such models (c.f. [Zhang et al., 2017; Kulkarni et al., 2019] and reasoning with uncertainty over them [Sreedharan et al., 2018]. It is true that this estimate might be different from the ground truth. However, an agent can only plan and explain with what it knows.}
Authors in [Chakraborti et al., 2017b] explored many such solutions—including ones that minimize length, called **minimally complete explanations** or MCEs. However, this was done post facto, i.e., the plan was already generated and it was just a matter of finding the best explanation for it. This not only ignores the possibility of finding better plans that are also optimal but with smaller explanations, but also misses avenues of compromise whereby the planner sacrifices its optimality to reduce the overhead of the explanation process.

### 3 The MEGA Algorithm

We bring the notions of explicability and explanations together in a novel planning technique MEGA (Multi-model Explanation Generation Algorithm) that trades off the relative cost of explicability to providing explanations during the plan generation process itself\(^2\). The output of MEGA is a plan \(\pi\) and an explanation \(E\) such that (1) \(\pi\) is executable in the robot’s model, and with the explanation (2) in the form of model updates it is (3) optimal in the updated human model while (4) the cost (length) of the explanations and the cost of deviation from optimality in its own model to be explicable is traded off according to a constant \(\alpha\) –

1. \(\delta_{\mathcal{M}^R}(\mathcal{G}^R, \pi) = |\mathcal{G}^R|\)
2. \(\bar{\mathcal{M}}^R \leftarrow \mathcal{M}^R + E\)
3. \(C(\pi, \bar{\mathcal{M}}^R) = \bar{c}_{\mathcal{M}^R}\) ; and
4. \(\pi = \arg \min_\pi \{ |E| + \alpha \times |C(\pi, \mathcal{M}^R) - \bar{c}_{\mathcal{M}^R}| \} \).

The objective thus takes into account the cost of choosing a particular plan by considering the cost difference (distance) with the optimal plan and the cost to explain it. The trade-off is thus not with respect to the total cost of the generated plan but the additional cost it suffers (but can avoid) in order to appear explicable. Clearly, with higher values of \(\alpha\) the planner will produce plans that require more explanation; with lower \(\alpha\) it will generate more explicable plans. The cost of an explanation not only includes the cognitive burden on the human in understanding/processing it but also the cost of communicating it from the point of view of the robot. For the purposes of this paper, we use explanation length as a proxy for both aspects of explanation costs. For example, the larger an explanation, the harder it may be to understand for the human – existing work [Chakraborti et al., 2017b] also use the same assumption. Similarly, a robot in a collapsed building during a search and rescue operation, or a rover on Mars, may have limited bandwidth for communication and prefer shorter explanations. \(\alpha\) thus has to be determined by the designer. As we show later in the evaluations, the decision of \(\alpha\) should also be based on the target population and the choice may not be static – i.e., the robot can vary it depending on its situation (e.g. if it is able to communicate more).

In the illustrative examples of the robot in the USAR task, the first plan it came up with (involving a slightly suboptimal plan and a short explanation) was indeed for lower value of \(\alpha\)

\(^2\)As in [Chakraborti et al., 2017b] we assume that the human mental model is known and has the same computation power ([Chakraborti et al., 2017b] also suggests possible ways to address these issues, the same discussions apply here as well).
Algorithm 1 MEGA

1: procedure MEGA-SEARCH
2:  
3: \quad \text{Input: HAP } \mathcal{W} = (\mathcal{M}^R, \mathcal{M}_R^d), \alpha
4: \quad \text{Output: Plan } \pi \text{ and Explanation } \mathcal{E}
5: fringe \leftarrow \text{PriorityQueue}() \triangleright \text{Closed list}
6: c_list \leftarrow \{\} \triangleright \text{Track node with min. value of obj.}
7: fringe.push((\mathcal{M}_R^d, \{\}, \text{priority} = 0)
8: while True do
9:   \langle \mathcal{M}, \mathcal{E} \rangle, c \leftarrow \text{fringe.pop}(\mathcal{M})
10:   \text{if OBJ_VAL(⟨M, E⟩) \leq OBJ_VAL(N_{min}) then}
11:      N_{min} \leftarrow \langle \mathcal{M}, \mathcal{E} \rangle \triangleright \text{Update min node}
12:   end if
13:   for \forall \pi^*_\mathcal{M} do \triangleright \text{This is relaxed in optimistic version}
14:      if C(\pi^*_\mathcal{M}, \mathcal{M}^R) = C^*_\mathcal{M}^R then
15:         \triangleright \text{Search is complete when } \pi^*_\mathcal{M} \text{ is optimal in } \mathcal{M}^R
16:         \langle \mathcal{M}_{min}, \mathcal{E}_{min} \rangle \leftarrow N_{min}
17:         return (\pi_{M_{min}}, \mathcal{E}_{min})
18:      else
19:         c_list \leftarrow c_list \cup \mathcal{M}
20:         for f \in \Gamma(\mathcal{M}) \setminus \Gamma(\mathcal{M}^R) do
21:            \triangleright \text{Misconceptions in the mental model}
22:            \lambda \leftarrow \{1, \{\mathcal{M}\}, \{f\} \triangleright \text{Remove from } \mathcal{M}
23:            if \delta_{\mathcal{M}_R^d, \mathcal{M}}(\Gamma(\mathcal{M}), \lambda) \not\in c_list then
24:               \triangleright \text{fringe.push}(\delta_{\mathcal{M}_R^d, \mathcal{M}}(\Gamma(\mathcal{M}), \lambda)
25:               \mathcal{E} \cup \lambda, c + 1)
26:          end if
27:         end for
28:         for f \in \Gamma(\mathcal{M}^R) \setminus \Gamma(\mathcal{M}) do
29:            \triangleright \text{Missing conditions in the mental model}
30:            \lambda \leftarrow \{1, \{\mathcal{M}\}, \{f\}, \{\} \triangleright \text{Add to } \mathcal{M}
31:            if \delta_{\mathcal{M}_R^d, \mathcal{M}}(\Gamma(\mathcal{M}), \lambda) \not\in c_list then
32:               \triangleright \text{fringe.push}(\delta_{\mathcal{M}_R^d, \mathcal{M}}(\Gamma(\mathcal{M}), \lambda)
33:               \mathcal{E} \cup \lambda, c + 1)
34:         end if
35:      end for
36:   end if
37: end while
38: 
39: procedure OBJ_VAL(⟨\mathcal{M}, \mathcal{E}⟩)
40:  \text{return } |\mathcal{E}| + \alpha \times \min_{\pi^*_\mathcal{M}} C(\pi^*_\mathcal{M}, \mathcal{M}^R) - C^*_\mathcal{M}^R
41: end procedure

the penalty is exactly zero). In general this works for any \( \alpha \geq |MCE| \) but since an MCE will only be known retrospectively after the search is complete, the above condition suffices since the entire model difference is known up front and is the largest possible explanation in the worst case.

**Property 3** \( \alpha = 0 \) yields the most explicable plan.

Under this condition, the planner minimizes the cost of explanations only – i.e. it will produce the plan that requires the shortest explanation, and hence the most explicable plan. Note that this is distinct from just computing the optimal plan in the mental model, since such a plan may not be executable in the robot model so that some explanations are required even in the worst case. This is also a welcome addition to the “explicability only” view of planning in [Zhang et al., 2017; Kulkarni et al., 2019] which do not deal with situations where a completely explicable plan does not exist, as done here using the explanations associated with the generated plans.

**Property 4** MEGA-search is required only once per problem, and is independent of \( \alpha \).

The algorithm terminates only after all the nodes containing a minimally complete explanation have been explored. This means that for different values of \( \alpha \), the agent only needs to post-process the nodes with the new objective function in mind. Thus, a large part of the reasoning process for a particular problem can be pre-computed.

**Property 5** A balanced solution is non-unique.

This is similar to standalone explicable plans and plan explanations – i.e. there can be many solutions to choose from, for a given \( \alpha \). Interestingly, solutions that are equally good according to the cost model can turn out to be different in usefulness to the human, as investigated recently in [Zahedi et al., 2019] in the context of plan explanations only. This can have similar implications to balanced solutions as well.

**Approximate MEGA**

MEGA evaluates executability (in the robot model) of all optimal plans within each intermediate model during search. This is quite expensive. Instead, we implement MEGA-approx that does this check only for the first optimal plan that gets computed. This means that, in Alogrithm 1, we drop the loop (line 16) and have a single optimality cost (line 38). This has the following consequence.

**Property 6** MEGA-approx is not complete.

MEGA-approx is an optimistic version of MEGA and is not guaranteed to find all balanced solutions. This is because in each search node we are checking for whether an optimal plans – the first one that gets computed – is executable in the robot model, and moving on if not. In models where multiple optimal plans are possible, and some are executable in the robot model while others are not, this will result in MEGA-approx discarding certain models as viable solutions where a balanced plan was actually possible. The resulting incompleteness of the search means we lose Properties 1 and, but it also allows us to compare directly to [Chakraborti et al., 2017b] where the optimal plan is fixed.

## 4 Empirical Evaluations

We will now provide evaluations of MEGA-approx demonstrating the trade-off in the cost and computation time of plans with respect to varying size of the model difference and the hyper-parameter \( \alpha \). We will then report on human factor studies on how this trade-off is received by users. The former evaluates from the perspective of the robot which is able to minimize communication but also the penalty due to explicability. The user study instead evaluates the effect of this on the human. The code is available at https://bit.ly/2XTKHz0.
4.1 Part-1: Cost Trade-off

\(\alpha\) determines how much an agent is willing to sacrifice optimality versus the explanation cost. We will illustrate this trade-off on modified versions of two popular IPC domains.

**The Rover (Meets a Martian) Domain**

Here the IPC Mars Rover has undergone an update whereby it can carry the rock and soil samples needed for a mission at the same time. This means that it does not need to empty the store before collecting new rock and soil samples anymore so that the new action definitions for sample_soil and sample_rock no longer contain the precondition (empty ?s).

During its mission it runs across a Martian who is unaware of the robot’s expanded storage capacity, and has an older, extremely cautious, model of the rover it has learned while spying on it from its cave. The Martian also believes that before the rover can perform take_image action, it needs to send the soil data and rock data of the waypoint from where it is taking the image. Clearly, if the rover was to follow this model, in order not to spook the Martian it will end up spending a lot of time performing unnecessary actions (like dropping old samples and collecting unnecessary samples). For example, if the rover is to communicate an image of an objective objective2, all it needs to do is move to a waypoint (waypoint3) from where objective2 is visible and perform the action –

\[
\text{(take_image waypoint3 objective2 camera0 high_res)}
\]

If the rover was to produce a plan that better represents the Martian’s expectations, it would look like –

\[
\begin{align*}
\text{(sample_soil store waypoint3)} \\
\text{(communicate_soil_data waypoint3 waypoint3 waypoint0)} \\
\text{(drop_off store)} \\
\text{(sample_rock store waypoint3)} \\
\text{(communicate_rock_data waypoint3 waypoint3 waypoint0)} \\
\text{(take_image waypoint3 objective1 camera0 high_res)}
\end{align*}
\]

If the rover uses an MCE here, it ends up explaining 6 model differences. In some cases, this may be acceptable, but in others, it may make more sense for the rover to bear the extra cost rather than laboriously walk through all updates with an impatient Martian. Figure 4 shows how the explicability and explanation costs vary for problem instances in this domain. The algorithm converges to the smallest possible MCE, when \(\alpha\) is set to 1. For smaller \(\alpha\), MEGA saves explanation cost by choosing more explicable (and expensive) plans.

**The Barman (in a Bar) Domain**

Here, the brand new two-handed Barman robot is wowing onlookers with its single-handed skills, even as its admirers who may be unsure of its capabilities expect, much like the standard IPC domain, that it needs one hand free for actions like fill-shot, refill-shot, shake etc. This means that to make a single shot of a cocktail with two shots of the same ingredient with three shots and one shaker, the human expects the robot to –

\[
\begin{align*}
\text{(fill-shot shot2 ingredient2 left right dispenser2)} \\
\text{(pour-shot-to-used-shaker shot2 ingredient3 shaker1 left)} \\
\text{(refill-shot shot2 ingredient3 left right dispenser3)} \\
\text{(pour-shot-to-used-shaker shot2 ingredient3 shaker1 left)} \\
\text{(leave left shot2)} \\
\text{(grasp left shaker1)}
\end{align*}
\]

The robot can, however, directly start by picking both the shot and the shaker and does not need to put either of them down while making the cocktail. Similar to the Rover domain, we again illustrate (Figure 4) how at lower values of \(\alpha\) the robot generates plans that require less explanation. As \(\alpha\) increases the algorithm produces plans that require larger explanations with the explanations finally converging at the smallest MCE required for that problem.

| \(\Delta\) | \(|\mathcal{E}|\) | Time | \(|\mathcal{E}|\) | Time | \(|\mathcal{E}|\) | Time |
|----------|----------|------|----------|------|----------|------|
| \(\Delta = 2\) | | | \(\Delta = 7\) | | | \(\Delta = 10\) | |
| Rover | p1 | 0 | 1.22 | | 1 | 5.83 | | 3 | 143.84 |
| | p2 | 1 | 1.79 | | 5 | 125.64 | | 6 | 1061.82 |
| | p3 | 0 | 8.35 | | 2 | 10.46 | | 3 | 53.22 |
| Barman | p1 | 2 | 18.70 | | 6 | 163.94 | | 6 | 5576.06 |
| | p2 | 2 | 2.43 | | 4 | 57.83 | | 6 | 953.47 |
| | p3 | 2 | 45.32 | | 5 | 4183.55 | | 6 | 5061.50 |

Table 1: Runtime (secs) and size of explanations \(\mathcal{E}\) with respect to the size of model difference \(\Delta\).
Gains due to Trade-off
Table 1 illustrates how the length of explanations computed square off with the total model difference $\Delta$. Clearly, there are significant gains to be had in terms of minimality of explanations and the reduction in cost of explicable plans as a result of it. This is something the robot trades off internally by considering its limits of communication, cost model, etc. We will discuss the external effect of this (on the human) later in the discussion of human factors studies we conducted.

Computation Time
Contrary to classical notions of planning that occurs in state or plan space, we are now planning in the model space, i.e. every node in the search tree is a new planning problem. As seen in Table 1, this can be time consuming (even for the approximate version) with increasing number of model differences between the human and the robot, even as there are significant gains to be had in terms of minimality of explanations, and the reduction in cost of explicable plans as a result of it. MEGA-approx remains comparable with the original work on model reconciliation [Chakraborti et al., 2017b] which also employs model space search, though we are solving a harder problem (computing the plan in addition to its explanation). Interestingly, in contrast to [Chakraborti et al., 2017b], the time taken here (while still within the bounds of the IPC Optimal Track) is conceded at planning time rather than at explanation time, so the user does not have to actually ask for an explanation and wait.

An interested reader may also refer to existing works on model space search [Keren et al., 2016; Chakraborti et al., 2017b] which introduces heuristics and approximations which are equally applicable here and can considerably speed up the process. However, the focus of our work is instead on the interesting behaviors that emerge from considering explanations during the plan generation process.

4.2 Part-2: Human Factors Evaluations
We use the USAR domain introduced before to analyze how human subjects respond to the explicable versus explanations trade-off. The experimental setup (reproduced here in part of clarity) derives from those used to study the model reconciliation process in [Chakraborti et al., 2019b]. Here, we extend those results to balanced plans. Specifically, we set out to test two key hypotheses –

H1. Subjects would require explanations when the robot comes up with suboptimal plans.

H1a. Response to balanced plans should be indistinguishable from inexplicable / robot optimal plans.

H2. Subjects would require less explanations for explicable plans as opposed to balanced or robot optimal plans.

H1 is the key thesis of recent works on explanations [Chakraborti et al., 2017b; Sreedharan et al., 2018] that formulates the process of explanation as one of model reconciliation to achieve common grounds with respect to a plan’s optimality. This forms the basis of incorporating considerations of explanations in the plan generation process as well, as done in the paper, in the event of model differences with the human in the loop. H2 forms the other side of this coin and completes the motivation of computing balanced plans. Note that balanced plans would still appear suboptimal (and hence inexplicable) to the human even though they afford opportunities to the robot to explain less or perform a more optimal plan. Thus, we expect (H1a) their behavior to be identical in case of both robot optimal and balanced plans.

Experimental Setup
The experimental setup exposes the external commander’s interface to participants who get to analyze plans in a mock USAR scenario. The participants were incentivized to make sure that the explanation does indeed help them understand the optimality of the plans in question by formulating the interaction in the form of a game. 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. Figure 5 shows a screenshot of the interface which 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. A plan is illustrated in the form of a series of paths through various waypoints highlighted on the map. The participant had to identify if the plan shown is optimal. If unsure, they could ask for an explanation. The explanation was provided in the form of a set of changes to the player’s map. The player was awarded 50 points for correctly identifying the plan as either optimal or satisficing. Incorrect identification cost them 20 points. Every request for explanation further cost them 5 points, while skipping a map did not result in any penalty. Even though there were no incorrect plans in the dataset, the participants were told that selecting an inexecutable plan as either feasible or optimal would result in a penalty of 400 points, in order to deter them from guessing when they were unsure.

Each subject was paid $10 as compensation for their participation and received additional bonuses depending on how well the performed ($\leq 240$ to $\geq 540$ points). This was done to ensure that participants only ask for an explanation when they are unsure about the quality of the plan (due to small neg-
ative points on explanations) while they are also incentivized to identify the feasibility and optimality of the given plan correctly (large reward and penalty on doing this wrongly).

Each participant was shown 12 maps. For 6 of them, they were shown the optimal robot plan, and when they asked for an explanation, they were randomly shown different types of explanations from [Chakraborti et al., 2017b]. For the rest, they were either shown a (explicable) plan that is optimal in their model with no explanation or a balanced plan with a shorter explanation. We had 27 participants, 4 female and 22 male of age 19-31 (1 participant did not reveal their demographic) with a total of 382 responses across all maps.

Experimental Results

Figure 6 shows how people responded to different kinds of explanations / plans. These results are from the two problem instances that included both a balanced and a fully explicable plan. Out of 54 user responses to these, 13 were for explicable plans and 12 for the balanced ones. From the perspective of the human, the balanced plan and the robot optimal plan do not make any difference since both of them appear sub-optimal. This is evident from the fact that the click-through rate for explanations in these two conditions are similar (H1a) (the high click-through rates for perceived suboptimality conform to the expectations of H1a). Further, the rate of explanations is much less for explicable plans as desired (H2).

Table 2 shows the statistics of the explanations / plans. These results are from 124 problem instances that required MCEs as per [Chakraborti et al., 2017b], and 25 and 40 instances that contained balanced and explicable plans respectively. As desired, the robot gains in length of explanations but loses out in cost of plans produced as it progresses along the spectrum of optimal to explicable plans. Thus, while Table 2 demonstrates the explanation versus explicability trade-off from the robot’s point of view, Figure 6 shows how this trade-off is perceived from the human’s perspective.

It is interesting to see that in Figure 6 almost a third of the time participants still asked for explanations even when the plan was explicable, i.e. optimal in their map. This is an artifact of the risk-averse behavior incentivized by the gamification of the explanation process and indicative of the cognitive burden on the humans who are not (cost) optimal planners. Furthermore, the participants also did not ask for explanations around 20-25% of the time when they “should have” (i.e. suboptimal plan in the human model). There was no clear trend here (e.g. decreasing rate for explanations asked due to increasing trust) and was most likely due to limitations of inferential capability of humans. Thus, going forward, the objective function should look to incorporate the cost or difficulty of analyzing the plans and explanations from the point of view of the human in addition to that in MEGA(4) and Table 2 modeled from the perspective of the robot.

Finally, in Figure 7, we show how the participants responded to inexplicable plans, in terms of their click-through rate on the explanation request button. Figure 7(left) shows the % of times subjects asked for explanations while Figure 7(right) shows the same w.r.t. the number of participants. They indicate the variance of human response to the explicability-explanations trade-off. Such information can be used to model the $\alpha$ parameter to situate the explicability versus explanation trade-off according to preferences of individual users. It is interesting to see that the distribution of participants (right inset) seem to be bimodal indicating that subjects are either particularly skewed towards risk-averse behavior or not, rather than a normal distribution of responses to the explanation-explicability trade-off. This is somewhat counter-intuitive and against expectations (H1) and further motivates the need for learning $\alpha$ interactively.

5 Conclusion

We saw how an agent can be human-aware by balancing the cost of departure from optimality (in order to conform to human expectations) versus the cost of explaining away causes of perceived suboptimality. It is well known how humans make better decisions when they have to explain [Mercier and Sperber, 2011]. In this work, in being able to reason about the explainability of its decisions, an AI planning agent is similarly able to make better decisions by explicitly considering the implications of its behavior on the human mental model. The work leaves open several intriguing avenues of further research, including how an agent can consider implicit communication of model differences via ontic and epistemic effects of its actions. In ongoing work [Sreedharan et al., 2019], we are exploring these ideas in a unified framework.

Acknowledgments

Majority of this work was done when all the authors were at Arizona State University. This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, and N00014-18-1-2840, AFOSR grant FA9550-18-1-0067 and NASA grant NNX17AD06G.
References


