Balancing Explicability and Explanations for Human-Aware Planning

Tathagata Chakraborti[1*], **Sarath Sreedharan**[2*], **Subbarao Kambhampati**[2]

[1]IBM Research AI, Cambridge MA 02142 USA
[2]Arizona 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 a planning agent while interacting with a human to use, in the process of its deliberation, not only its own model $\mathcal{M}^R$ of the task, but also the model $\mathcal{M}_h^R$ that the human thinks it has. This mental model [Chakraborti *et al.*, 2017a] is in addition to the task model of the human M_r^H (denoting their beliefs, intentions and capabilities). This is, in essence, the fundamental thesis of the recent works on **plan explanations** [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 the robot's belief of it) are involved in the planning process. The need for explicable planning or plan explanations occur when these two models – $\mathcal{M}^R$ and $\mathcal{M}_h^R$ – diverge. This means that the optimal plans in the respective models – $\pi^*_{\mathcal{M}^R}$ and $\pi^*_{\mathcal{M}_h^R}$ – may not be the same and hence optimal behavior of the robot in its own model seems inexplicable to the human.

– In **explicable planning** [Zhang *et al.*, 2017; Dragan *et al.*, 2013], the robot produces a plan $\bar{\pi}$ that is closer to the human's expected plan, i.e. $\bar{\pi} \approx \pi^*_{\mathcal{M}_h^R}$.

– During **plan explanation** [Chakraborti *et al.*, 2017b; Rosenthal *et al.*, 2016], it attempts to update the human's mental model to an intermediate model $\bar{\mathcal{M}}_h^R$ 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. $\pi^*_{\mathcal{M}^R} \equiv \pi^*_{\bar{\mathcal{M}}_h^R}$.

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 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 plan generation process itself.

– **[1a]** We show how, unlike existing explanation-only algorithms, this allows us to generate explanations that are 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 algorithms, we can deal with situations where an explicable plan does not exist by being able to reconcile model differences in the same planning framework.

[*]Equal contribution.

[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.

Illustrative Example We start with an illustration of the explicabilty-explanation trade-off in a typical Urban Search And Reconnaissance (USAR) task [Bartlett, 2015]. *A video can be viewed at (https://youtu.be/Yzp4FU6Vn0M).* Here a remote robot is performing a search controlled partly or fully by an external human commander. The external has a map of the environment, but this map is no longer accurate – e.g. new paths may have opened up, or older paths may no longer be available, due to rubble. 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. Using our algorithm, the robot deals with these model differences by modulating how explicable its plans need to be for the external supervisor versus how much explanation it might need to provide for its plans. In the first part of the demonstration, the robot chooses a plan that requires the least amount of explanation, i.e. the most explicable plan. The robot only needed to explain a single model change to make its plan optimal. It later uses the same algorithm to switch to its optimal plan with the help of a larger explanation which conveys the optimality of the current plan as well as the infeasibility of the human's expected plan.

Related Work Efforts to make planning more "human-aware" have largely focused on incorporating an agent's understanding of the human model $\mathcal{M}^H$ into its decision making process. Since then the importance of considering the human's understanding $\mathcal{M}_h^R$ of the agent's actual model $\mathcal{M}^R$ 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]. The need for such agents to be able to explain their behavior has also been emphasized [Chakraborti *et al.*, 2017b; Langley *et al.*, 2017; Fox *et al.*, 2017]. These considerations allow a human-aware agent to conceive novel and interesting behaviors both in the space of plans and models. For example, in the model space, the modifications to the human mental model $\mathcal{M}_h^R$ is used for explanations in [Chakraborti *et al.*, 2017b] while reasoning over the actual model $\mathcal{M}^H$ can reveal interesting behavior by affecting the belief state of the human, such as in planning for serendipity [Chakraborti *et al.*, 2015]. In the plan space, a human-aware agent can use $\mathcal{M}^H$ and $\mathcal{M}_h^R$ 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; Tomic *et al.*, 2014] and expectations [Dragan *et al.*, 2013; Zhang *et al.*, 2017; Kulkarni *et al.*, 2019] and create plans that help the human understand the robots 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.*, 2019].

2 Human-Aware Planning

A Classical Planning Problem is a tuple $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}, \mathcal{G} \rangle$ with domain $\mathcal{D} = \langle F, A \rangle$ - where F is a set of fluents that define a state $s \subseteq F$, and A is a set of actions - and initial and goal states $\mathcal{I}, \mathcal{G} \subseteq F$. Action $a \in A$ is a tuple $\langle c_a, \mathit{pre}(a), \mathit{eff}^{\pm}(a) \rangle$ where c_a is the cost, and $\mathit{pre}(a), \mathit{eff}^{\pm}(a) \subseteq F$ are the preconditions and add/delete effects, i.e. $\delta_{\mathcal{M}}(s, a) \models \perp$ *if* $s \not\models \mathit{pre}(a)$; *else* $\delta_{\mathcal{M}}(s, a) \models s \cup \mathit{eff}^{+}(a) \setminus \mathit{eff}^{-}(a)$ where $\delta_{\mathcal{M}}(\cdot)$ is the transition function.

Note that the "model" $\mathcal{M}$ of a planning problem includes the action model *as well as the initial and goal states of an agent*. The solution to $\mathcal{M}$ is a sequence of actions or a (satisficing) *plan* $\pi = \langle a_1, a_2, \ldots, a_n \rangle$ such that $\delta_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$. The cost of a plan π is $C(\pi, \mathcal{M}) = \sum_{a \in \pi} c_a$ if $\delta_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$; ∞ otherwise. The optimal plan has cost $C^*_{\mathcal{M}}$.

A Human-Aware Planning (HAP) Problem is the tuple $\Psi = \langle \mathcal{M}^R, \mathcal{M}_h^R \rangle$ where $\mathcal{M}^R = \langle D^R, \mathcal{I}^R, \mathcal{G}^R \rangle$ and $\mathcal{M}_h^R = \langle D_h^R, \mathcal{I}_h^R, \mathcal{G}_h^R \rangle$ are the planner's model of a planning problem and the human's understanding of the same[1]. There can be two kinds of solutions to HAP problems, as discussed below.

[1] Explicable Plans An explicable solution to HAP is a plan π (1) executable in the robot's model and (2) closest to the expected (optimal) plan in the human's model –

(1) $\delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \models \mathcal{G}^R$; and

(2) $C(\pi, \mathcal{M}_h^R) \approx C^*_{\mathcal{M}_h^R}$.

"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] 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 human's updated model of the problem. Thus, a solution involves a plan π and an explanation $\mathcal{E}$ such that –

[1]Note that this **does not** assume that humans use an explicitly represented symbolic domain to plan. The robot only uses this to represent the possible information content of that model. It, of course, does not have direct access to it but uses whatever estimate it has. 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.*, 2018a; Sreedharan *et al.*, 2018b]. 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.

(1) $C(\pi, \mathcal{M}^R) = C^*_{\mathcal{M}^R}$;

(2) $\bar{\mathcal{M}}^R_h \longleftarrow \mathcal{M}^R_h + \mathcal{E}$; and

(3) $C(\pi, \bar{\mathcal{M}}^R_h) = C^*_{\bar{\mathcal{M}}^R_h}$.

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.

[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 optimal 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 equally optimal) with smaller explanations, but also misses avenues of compromise whereby the planner sacrifices its optimality to further reduce overhead in 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 π and an explanation $\mathcal{E}$ such that (1) π 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 α –

(1) $\delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \models \mathcal{G}^R$;

(2) $\bar{\mathcal{M}}^R_h \longleftarrow \mathcal{M}^R_h + \mathcal{E}$;

(3) $C(\pi, \bar{\mathcal{M}}^R_h) = C^*_{\bar{\mathcal{M}}^R_h}$; and

(4) $\pi = \arg\min_\pi \{ \; |\mathcal{E}| \; + \; \alpha \times |\, C(\pi, \mathcal{M}^R) - C^*_{\mathcal{M}^R} \,| \; \}$.

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 α the planner will produce plans that require more explanation; with lower α 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. α thus has to be determined by the designer. As we show later in the evaluations, the decision of α 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 α while the second one (optimal with a larger explanation) was for a higher value of α. Interestingly, the first case is also 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.

Model Space Search We employ a *model space* A^* search to compute the expected plan and explanations for a given value of α. Similar to [Chakraborti *et al.*, 2017b] we define a state representation over planning problems with a mapping function $\Gamma : \; {}^{a}\mathcal{M} \mapsto \mathcal{F}$ which represents a planning problem by transforming every condition in it into a predicate. The set Λ of actions contains unit model change actions which make a single change to a domain at a time. The algorithm starts by initializing the min node tuple ($\mathcal{N}$) with the human mental model M^R_h and an empty explanation. For each new possible model $\bar{M}$ generated during model space search, we test if the objective value of the new node is smaller than the current min node. We stop the search once we identify a model that is capable of producing a plan that is also optimal in the robot's own model. This is different from the original MCE-search [Chakraborti *et al.*, 2017b] where the authors are trying to find the *first* node where a given plan is optimal. Finally, we select the node with the best objective value as the solution.

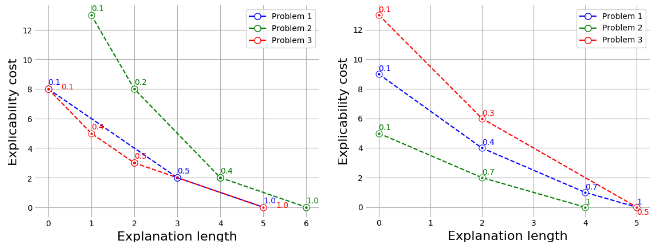
Property 1 MEGA yields the smallest possible explanation for a given human-aware planning problem.

This means that with a high enough α the algorithm is guaranteed to compute the best possible plan for the planner as well as the smallest explanation associated with it. This is by construction of the search process itself, i.e. the search only terminates after the all the nodes that allow $C(\pi, \bar{\mathcal{M}}^R_h) = C^*_{\bar{\mathcal{M}}^R_h}$ have been exhausted. *This is beyond what is offered by the model reconciliation search in [Chakraborti* et al.*, 2017b], which only computes the smallest explanation* given *a plan that is optimal in the planner's model.*

Property 2 $\alpha = |\, \mathcal{M}^R \; \Delta \; \mathcal{M}^R_h \,|$ (i.e. the total number of differences in the models) yields the most optimal plan in the planner's model along with the minimal explanation possible given a human-aware planning problem.

This is easy to see, since with $\forall \mathcal{E}, \; |\mathcal{E}| \leq |\, \mathcal{M}^R \; \Delta \; \mathcal{M}^R_h \,|$, the latter being the total model difference, the penalty for de-

[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).



(a) The Rover Domain (b) The Barman Domain

Figure 1: Explicability vs. explanation costs w.r.t. α.

parture from explicable plans is high enough that the planner must choose from possible explanations only (note that the explicability penalty is always positive until the search hits the nodes with $C(\pi, \bar{\mathcal{M}}_h^R) = C^*_{\bar{\mathcal{M}}_h^R}$, at which point onwards 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 human's model, since such a plan may not be executable in the planner's model so that some explanations are required even in the worst case. This is also a welcome addition to the "explicability only" view of plan generation introduced in [Zhang *et al.*, 2017; Kulkarni *et al.*, 2019] which cannot 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 α.

The algorithm terminates only after all the nodes containing a minimally complete explanation have been explored. This means that for different values of α, 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.

Also, note that, similar to standalone explicable plans and plan explanations, a balanced solution is non-unique – i.e. there can be many solutions to choose from, for a given α. Interestingly, solutions that are equally good according to the cost model can turn out to be different in usefulness to the human in the loop, as investigated recently in [Zahedi *et al.*, 2019] in the context of plan explanations only, and can have similar implications to balanced solutions as well.

4 Empirical Evaluations

We will now provide internal evaluations of `MEGA` in modified versions of two well-known IPC domains `Rover` and `Barman` [International Planning Competition, 2011] demonstrating the trade-off in the cost and computation time of plans

Domain Name	Problem	$\Delta = 2$		$\Delta = 7$		$\Delta = 10$	
		$\lvert\mathcal{E}\rvert$	Time (secs)	$\lvert\mathcal{E}\rvert$	Time (secs)	$\lvert\mathcal{E}\rvert$	Time (secs)
`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 and size of explanations w.r.t. size of model difference.

with respect to varying size of the model difference and the hyper-parameter α. We will then report on human factor studies on how this trade-off is received by users. Note that the two flavors of evaluations are done with different motivations. 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 already available online and will be linked here after double-blind review.*

4.1 Part-1: Cost Trade-off

α determines how much an agent is willing to sacrifice optimality versus the cost of explanation. We will now illustrate this trade-off on modified versions of two well-known IPC [International Planning Competition, 2011] 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. It believes that any time the Rover collects a rock sample, it also needs to collect a soil sample and need to communicate this information to the lander. 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 –

```
(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 –

```
(sample_soil store waypoint3)
(communicate_soil_data general waypoint3 waypoint3 waypoint0)
(drop_off store)
(sample_rock store waypoint3)
(communicate_rock_data general waypoint3 waypoint3 waypoint0)
(take_image waypoint3 objective1 camera0 high_res)
```

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 1 shows how the explicability and explanation costs vary for problem instances in this domain. The algorithm converges to the smallest possible MCE, when α is set to 1. For smaller α, 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 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 –

```
(fill-shot shot2 ingredient2 left right dispenser2)
(pour-shot-to-used-shaker shot2 ingredient3 shaker1 left)
(refill-shot shot2 ingredient3 left right dispenser3)
(pour-shot-to-used-shaker shot2 ingredient3 shaker1 left)
(leave left shot2)
(grasp left shaker1)
```

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 1) how at lower values of α the robot generates plans that require less explanation. As α increases the algorithm produces plans that require larger explanations with the explanations finally converging at the smallest MCE required for that problem.

Gains due to Trade-off Table 1 illustrate how the length of explanations computed square off with the total model difference Δ. 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. As mentioned before, 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 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. It is also 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.*

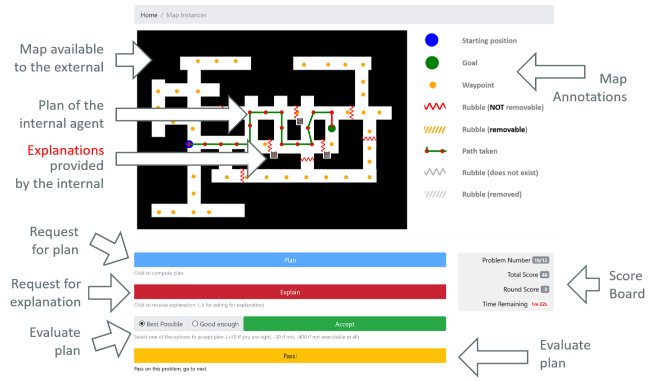


Figure 2: Interface to the external commander in the study.

4.2 Part-2: Human Factors Evaluations

We use the USAR domain to analyze how human subjects respond to the explicability versus explanations trade-off. Specifically, we set out to test two key hypothesis –

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.*, 2018a] 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 2 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 negative 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 was 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 3 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 suboptimal. 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 3 shows how this trade-off is perceived from the human's perspective.

It is interesting to see that in Figure 3 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.

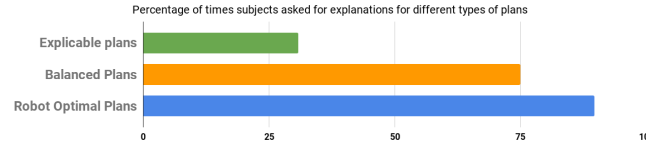


Figure 3: Reduced request for explanations to explicable plans.

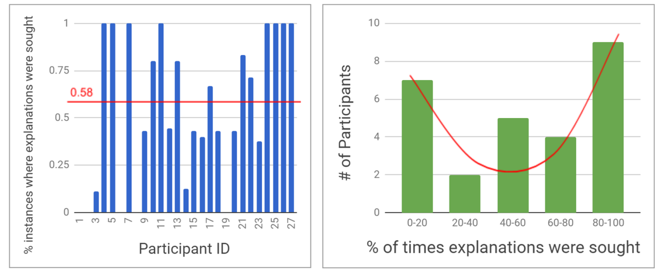


Figure 4: Click-through rates for explanations.

Optimal Plan		Balanced Plan		Explicable Plan	
$\lvert\mathcal{E}\rvert$	$C(\pi, \mathcal{M}^R)$	$\lvert\mathcal{E}\rvert$	$C(\pi, \mathcal{M}^R)$	$\lvert\mathcal{E}\rvert$	$C(\pi, \mathcal{M}^R)$
2.5	5.5	1	8.5	-	16

Table 2: Statistics of explicability vs. explanation trade-off.

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 4, we show how the participants responded to inexplicable plans, in terms of their click-through rate on the explanation request button. Figure 4(left) shows the % of times subjects asked for explanations while Figure 4(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 α 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 α interactively.

Conclusion

In conclusion, we saw how an agent can achieve human-aware behavior while keeping in mind the cost of departure from its own optimality which could be explained away. We showed how this can be achieved by a novel model-space search algorithm and evaluated different properties of the approach on well-known IPC domains as well as through human factors experiments in a mock USAR domain. This raises 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.

Acknowledgments
Some 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

[Alami *et al.*, 2006] Rachid Alami, Aurélie Clodic, Vincent Montreuil, Emrah Akin Sisbot, and Raja Chatila. Toward Human-Aware Robot Task Planning. In *AAAI Spring Symposium*, 2006.

[Alami *et al.*, 2014] Rachid Alami, Mamoun Gharbi, Benjamin Vadant, Raphaël Lallement, and Adolfo Suarez. On human-aware task and motion planning abilities for a teammate robot. In *Human-Robot Collaboration for Industrial Manufacturing Workshop, RSS*, 2014.

[Bartlett, 2015] Cade Earl Bartlett. Communication between Teammates in Urban Search and Rescue. *Thesis*, 2015.

[Chakraborti *et al.*, 2015] Tathagata Chakraborti, Gordon Briggs, Kartik Talamadupula, Yu Zhang, Matthias Scheutz, David E. Smith, and Subbarao Kambhampati. Planning for serendipity. In *IROS*, 2015.

[Chakraborti *et al.*, 2017a] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. AI Challenges in Human-Robot Cognitive Teaming. *arXiv preprint arXiv:1707.04775*, 2017.

[Chakraborti *et al.*, 2017b] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*, 2017.

[Chakraborti *et al.*, 2019] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *ICAPS*, 2019.

[Cirillo *et al.*, 2010] Marcello Cirillo, Lars Karlsson, and Alessandro Saffiotti. Human-aware Task Planning: An Application to Mobile Robots. *Transactions on Intelligent Systems and Technology (TIST)*, 2010.

[Dragan *et al.*, 2013] Anca Dragan, Kenton Lee, and Siddhartha Srinivasa. Legibility and predictability of robot motion. In *HRI*, 2013.

[Fox *et al.*, 2017] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable Planning. In *IJCAI Workshop on Explainable AI (XAI)*, 2017.

[International Planning Competition, 2011] International Planning Competition. IPC Competition Domains. https://goo.gl/i35bxc, 2011.

[Keren *et al.*, 2016] Sarah Keren, Avigdor Gal, and Erez Karpas. Privacy Preserving Plans in Partially Observable Environments. In *IJCAI*, 2016.

[Koeckemann *et al.*, 2014] Uwe Koeckemann, Federico Pecora, and Lars Karlsson. Grandpa Hates Robots - Interaction Constraints for Planning in Inhabited Environments. In *AAAI*, 2014.

[Kulkarni *et al.*, 2019] Anagha Kulkarni, Yantian Zha, Tathagata Chakraborti, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. Explicable Robot Planning as Minimizing Distance from Expected Behavior. *AAMAS EA*, 2019.

[Langley *et al.*, 2017] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *AAAI/IAAI*, 2017.

[Rosenthal *et al.*, 2016] Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. Verbalization: Narration of autonomous robot experience. In *IJCAI*, 2016.

[Sadigh *et al.*, 2016] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, 2016.

[Sreedharan *et al.*, 2018a] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Handling Model Uncertainty and Multiplicity in Explanations as Model Reconciliation. In *ICAPS*, 2018.

[Sreedharan *et al.*, 2018b] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. Hierarchical expertise level modeling for user specific contrastive explanations. In *IJCAI*, pages 4829–4836, 2018.

[Talamadupula *et al.*, 2014] Kartik Talamadupula, Gordon Briggs, Tathagata Chakraborti, Matthias Scheutz, and Subbarao Kambhampati. Coordination in human-robot teams using mental modeling and plan recognition. In *IROS*, 2014.

[Tomic *et al.*, 2014] Stevan Tomic, Federico Pecora, and Alessandro Saffiotti. Too Cool for School??? Adding Social Constraints in Human Aware Planning. In *Workshop on Cognitive Robotics (CogRob)*, 2014.

[Zahedi *et al.*, 2019] Zahra Zahedi, Alberto Olmo, Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. Towards Understanding User Preferences for Explanation Types in Explanation as Model Reconciliation. In *HRI Late Breaking Report*, 2019.

[Zhang *et al.*, 2017] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan Explicability and Predictability for Robot Task Planning. In *ICRA*, 2017.