Human-Aware Planning Revisited: *A Tale of Three Models*

Tathagata Chakraborti and **Sarath Sreedharan** and **Subbarao Kambhampati**
Arizona State University, Tempe, AZ 85281 USA
{ tchakra2, ssreedh3, rao } @ asu.edu

Abstract

Human-aware planning requires an agent to be aware of the mental model of the humans, in addition to their physical or capability model. This not only allows an agent to envisage the desired roles of the human in a joint plan but also anticipate how its plan will be *perceived* by the latter. The human mental model becomes especially useful in the context of an explainable planning (XAIP) agent since an explanatory process cannot be a soliloquy, i.e. it must incorporate the human's beliefs and expectations of the planner. In this paper, we survey our recent efforts in this direction.

Cognitive AI teaming (Chakraborti *et al.* 2017a) requires a planner to perform argumentation over a set of models during the plan generation process. This is illustrated in Figure 1. Here, $\mathcal{M}^R$ is the model of the agent embodying the planner (e.g. a robot), and $\mathcal{M}^H$ is the model of the human in the loop. Further, $\mathcal{M}^R_h$ is the model the human thinks the robot has, and $\mathcal{M}^H_r$ is the model that the robot thinks the human has. Finally, $\tilde{\mathcal{M}}^R_h$ is the robot's approximation of $\mathcal{M}^R_h$; for the rest of the paper we will be using $\mathcal{M}^R_h$ to refer to both since, for all intents and purposes, this is all the robot has access to. Note that the *human mental model* $\mathcal{M}^R_h$ is in addition to the (robot's belief of the) *human model* $\mathcal{M}^H_r$ traditionally encountered in human-robot teaming (HRT) settings and is, in essence, the fundamental thesis of the recent works on *plan explanations* (Chakraborti *et al.* 2017b) and *explicable planning* (Zhang *et al.* 2017). The need for explicable planning or plan explanations occurs when the models – $\mathcal{M}^R$ and $\mathcal{M}^R_h$ – diverge so that the optimal plans in the respective models may not be the same and hence *optimal behavior of the robot in its own model is inexplicable to the human*. This is also true for discrepancies between $\mathcal{M}^H$ and $\mathcal{M}^H_r$ when the robot might reveal unrealistic expectations of the human in a joint plan.

An explainable planning (XAIP) agent (Fox *et al.* 2017; Langley *et al.* 2017; Weld and Bansal 2018) should be able to able to deal with such model differences and participate in explanatory dialog with the human such that both of them can be on the same page during a collaborative activity. This is referred to as *model reconciliation* (Chakraborti *et al.* 2017b) and forms the core of the explanatory process of an XAIP agent. In this paper, we look at the scope of problems engendered by this multi-model setting and describe

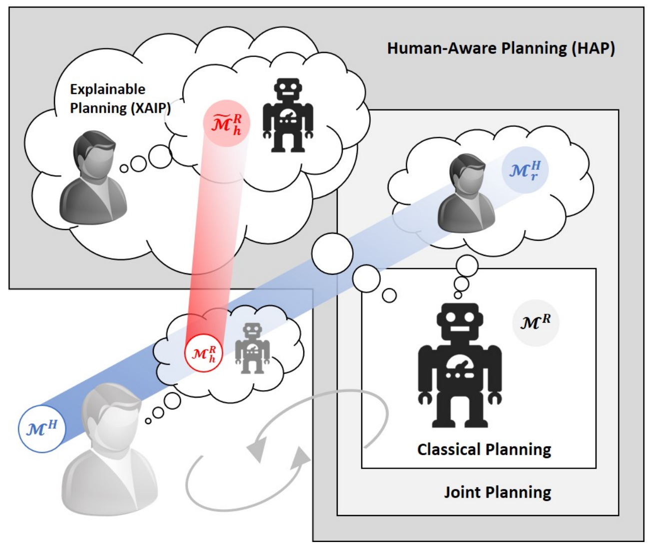

Figure 1: Argumentation over multiple models during the deliberative process of a human-aware planner (e.g. robot).

the recent work in this direction. Specifically –

- We outline the scope of behaviors engendered by human-aware planning, including joint planning as studied in teaming using the *human model*, as well as explicable planning with the *human mental model*;
- We situate the plan explanation problem in the context of *perceived* inexplicability of the robot's plans or behaviors due to differences in these models;
  - We discuss how the plan explanation process can be seen as one of *model reconciliation* where $\mathcal{M}^R_h$ (and/or $\mathcal{M}^H_r$) is brought closer to $\mathcal{M}^R$ ($\mathcal{M}^H$);
  - We discuss how explicability and explanation costs can be traded off during plan generation;
  - We discuss how this process can be adapted to handle uncertainty or multiple humans in the loop;
  - We discuss results of a user study that testify to the usefulness of the model reconciliation process;
- We point to ongoing work in the space of abstractions and deception using the human mental model.

Background

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 finite set of fluents that define a state $s \subseteq F$, and A is a finite 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. $\delta_{\mathcal{M}}(s, a) \models \bot$ *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. The cumulative transition function is given by $\delta_{\mathcal{M}}(s, \langle a_1, a_2, \ldots, a_n \rangle) = \delta_{\mathcal{M}}(\delta_{\mathcal{M}}(s, a_1), \langle a_2, \ldots, a_n \rangle)$.

This forms the classical definition of a planning problem (Russell and Norvig 2003) whose models are represented in the syntax of PDDL (McDermott *et al.* 1998). The solution to the planning problem 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 given by $C(\pi, \mathcal{M}) = \sum_{a \in \pi} c_a$ if $\delta_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$; ∞ otherwise. The cheapest plan $\pi^* = \arg\min_{\pi} C(\pi, \mathcal{M})$ is the (cost) optimal plan with cost $C^*_{\mathcal{M}}$.

In previous work (Nguyen *et al.* 2017) we introduced an updated representation of planning problems in the form of *annotated* models to account for uncertainty or incompleteness over the definition of a planning model. In addition to the standard preconditions and effects associated with actions, it introduces the notion of *possible* preconditions and effects which may or may not be realized in practice.

An Incomplete (Annotated) Model is the tuple $\mathbb{M} = \langle \mathbb{D}, \mathbb{I}, \mathbb{G} \rangle$ with a domain $\mathbb{D} = \langle F, \mathbb{A} \rangle$ – where F is a finite set of fluents that define a state $s \subseteq F$, and $\mathbb{A}$ is a finite set of annotated actions – and annotated initial and goal states $\mathbb{I} = \langle \mathcal{I}^0, \mathcal{I}^+ \rangle$, $\mathbb{G} = \langle \mathcal{G}^0, \mathcal{G}^+ \rangle$; $\mathcal{I}^0, \mathcal{G}^0, \mathcal{I}^+, \mathcal{G}^+ \subseteq F$. Action $a \in \mathbb{A}$ is a tuple $\langle c_a, \mathit{pre}(a), \widetilde{\mathit{pre}}(a), \mathit{eff}^{\pm}(a), \widetilde{\mathit{eff}}^{\pm}(a) \rangle$ where c_a is the cost and, in addition to its *known* preconditions and add/delete effects $\mathit{pre}(a), \mathit{eff}^{\pm}(a), \subseteq F$ each action also contains *possible preconditions* $\widetilde{\mathit{pre}}(a) \subseteq F$ containing propositions that action a *might* need as preconditions, and *possible add (delete) effects* $\widetilde{\mathit{eff}}^{\pm}(a) \subseteq F$) containing propositions that the action a *might* add (delete, respectively) after execution. Similarly, $\mathcal{I}^0, \mathcal{G}^0$ (and $\mathcal{I}^+, \mathcal{G}^+$) are the known (and possible) parts of the initial and goal states.

Each possible condition $f \in \widetilde{\mathit{pre}}(a) \cup \widetilde{\mathit{eff}}^{\pm}(a)$ has a probability $p(f)$ associated with it denoting how likely it is to appear as a known condition in the ground truth model – i.e. $p(f)$ measures the confidence with which that condition has been learned. The sets of known and possible conditions in $\mathcal{M}$ are called $\mathbb{S}_k(\Gamma(\mathcal{M}))$ and $\mathbb{S}_p(\Gamma(\mathcal{M}))$. Here Γ is a mapping function that converts domain model conditions into propositions in a meta space (Chakraborti *et al.* 2017b).

An *instantiation* of an annotated model $\mathbb{M}$ is a classical planning model where a subset of the possible conditions have been realized – given by the tuple $\mathit{inst}(\mathbb{M}) = \langle \mathcal{D}, \mathcal{I}, \mathcal{G} \rangle$ with domain $\mathcal{D} = \langle F, A \rangle$, initial and goal states $\mathcal{I} = \mathcal{I}^0 \cup \chi$; $\chi \subseteq \mathcal{I}^+$ and $\mathcal{G} = \mathcal{G}^0 \cup \chi$; $\chi \subseteq \mathcal{G}^+$ respectively, and action $A \ni a = \langle c_a, \mathit{pre}(a) \leftarrow \mathit{pre}(a) \cup \chi;\ \chi \subseteq \widetilde{\mathit{pre}}(a), \mathit{eff}^{\pm}(a) \leftarrow \mathit{eff}^{\pm}(a) \cup \chi;\ \chi \subseteq \widetilde{\mathit{eff}}^{\pm}(a) \rangle$. Given an annotated model with k possible conditions, there may be 2^k such instantiations, which forms its *completion set*.

The Likelihood $\mathcal{L}$ of an instantiation $\mathit{inst}(\mathbb{M})$ of the annotated model $\mathbb{M}$ is given by –

$$\mathcal{L}(\mathit{inst}(\mathbb{M})) = \prod_{f \in \mathbb{S}_p(\Gamma(\mathbb{M})) \wedge \mathbb{S}_k(\Gamma(\mathit{inst}(\mathbb{M})))} p(f) \times \prod_{f \in \mathbb{S}_p(\Gamma(\mathbb{M})) \setminus \mathbb{S}_k(\Gamma(\mathit{inst}(\mathbb{M})))} (1 - p(f))$$

Such models turn out to be especially useful for the representation and learning of human (mental) models from observations, where uncertainty after the learning process can be represented in terms of model annotations as in (Nguyen *et al.* 2017; Bryce *et al.* 2016). Let $\mathbb{M}^R_H$ be the culmination of a model learning process and $\{\mathcal{M}^R_{h_i}\}_i$ be the completion set of $\mathbb{M}^R_H$. Note that one of these models – $g(\mathbb{M}^R_H)$ – is the actual ground truth (i.e. the human's real mental model).

The USAR Domain

We will illustrate the algorithms in this paper in a typical (Bartlett 2015) 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. In general, differences in the models of the human and the robot can manifest in any form (e.g. the robot may have lost some capability or its goals may have changed). In the current setup, we deal with differences in the map of the environment as available to the human-robot team, i.e. these differences can then be compiled to differences only in the initial state of the planning problem (the human model has the original unaffected model of the world). This makes no difference to the proposed algorithms which treat all model changes equally.

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 aspects of task plans such as preconditions and effects in terms of rubble removal, collapsed halls, etc. and relevant abilities of the robot. As such, simulated USAR scenarios provide an ideal testbed for developing algorithms for effective human-robot interaction. Figure 2:B illustrates our setup (https://youtu.be/40Xol2GY7zE).

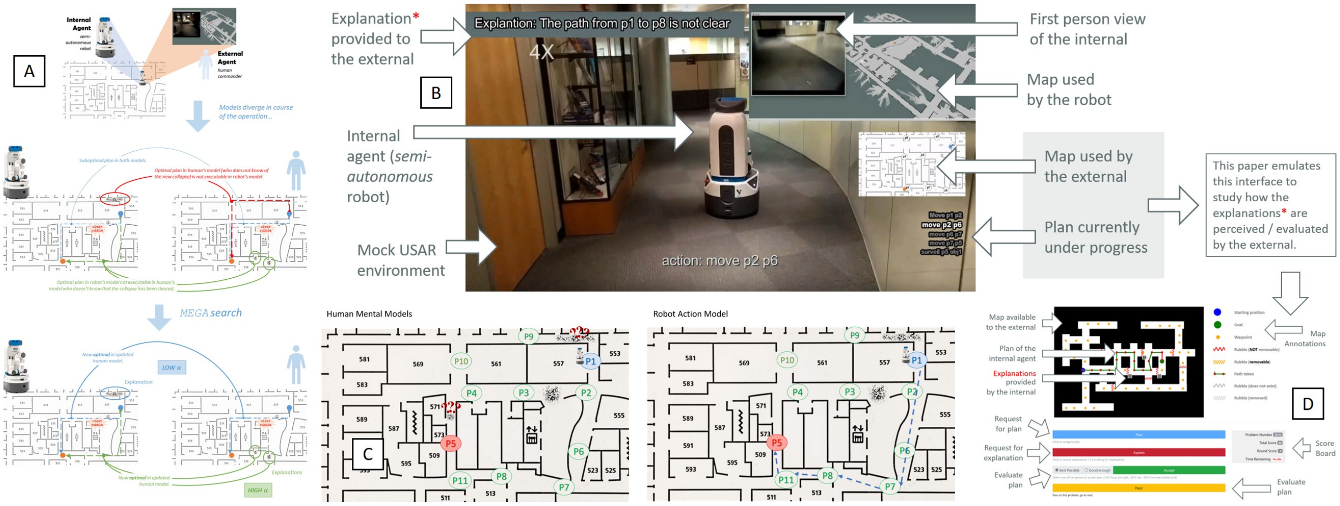

Figure 2: A mock USAR setting for studying the human-robot relationship in a typical disaster response team.

Human-Aware Planning Revisited

The human-aware planning paradigm introduces the *mental model* of the human in the loop into a planner's deliberative process, in addition to the planner's own model in the classical sense and the robot's estimate of the *human model*.

A Human-Aware Planning (HAP) Problem is given by the tuple $\Psi = \langle \mathcal{M}^R, \mathcal{M}^H_r, \mathcal{M}^R_h \rangle$ where $\mathcal{M}^R = \langle D^R, \mathcal{I}^R, \mathcal{G}^R \rangle$ is the planner's model of a planning problem, while $\mathcal{M}^R_h = \langle D^R_h, \mathcal{I}^R_h, \mathcal{G}^R_h \rangle$ is the human's understanding of the same, and $\mathcal{M}^H_r = \langle D^H_r, \mathcal{I}^H_r, \mathcal{G}^H_r \rangle$ is the planner's belief of the human's capability model.

The solution to the human-aware planning problem is a joint plan (Chakraborti *et al.* 2015) $\pi = \langle a_1, a_2, \ldots, a_n \rangle$; $a_i \in \{D^R \cup D^H_r\}$ such that $\delta_\Psi(\mathcal{I}^R \cup \mathcal{I}^H_r, \pi) \models \mathcal{G}^R \cup \mathcal{G}^H_r$. The robot's component in the plan is referred to as $\pi(R) = \langle a_i \mid a_i \in \pi \wedge D^R \rangle$; and similarly $\pi(H)$ for the human. Efforts to make planning more "human-aware" has largely focused on adapting $\pi(R)$ to meet the demands of $\pi(H)$ such as in (Alami *et al.* 2006; 2014; Cirillo *et al.* 2010; Koeckemann *et al.* 2014; Tomic *et al.* 2014; Cirillo 2010; Chakraborti *et al.* 2015; 2016; Talamadupula *et al.* 2014; Zhang *et al.* 2015) in the context of human-robot teams where a robot sacrifices optimality in its own model in favor of globally optimal joint plans. From the perspective of an XAIP agent, computation of the joint plan becomes more interesting when considering $\mathcal{M}^R_h$ as well, i.e. how $\pi(R)$ is *perceived* by the human. One solution is to be "explicable", i.e. make the robot conform to what the human expects of it.

Explicable Planning

An "explicable" solution to the human-aware planning problem is a plan π such that (1) it is executable (but may no longer be optimal) in the robot's model but is (2) "closer" to the expected plan in the human's model –

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

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

Such a plan is referred to as explicable because the human can explain it in their current mental model. "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 (Srivastava *et al.* 2007; Nguyen *et al.* 2012). In existing literature (Zhang *et al.* 2017; 2016; Kulkarni *et al.* 2016) 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.* 2016) from the mental model or *learned* (Zhang *et al.* 2017) through interactions in the form of affinity functions between plans and their purported goals. The solutions generated this way satisfy the planner's goal, as required by Condition (1), but are also biased towards the human's expectations as required by Condition (2) above.

It is interesting to note that, while mental modeling allows for human-awareness in the positive sense, it can also open up pathways for deception. Indeed, recent work (Kulkarni *et al.* 2018) has looked at how the concept of explicability can be flipped to obfuscate a robot's intentions.

Plan Explanations

The other approach would be to compute optimal plans in the planner's model (which may appear as inexplicable to the human) and provide an explanation of that plan in terms of the model differences – this is referred to as the process of *model reconciliation* (Chakraborti *et al.* 2017b). Although explanation of plans has been investigated in the past (c.f. (Kambhampati 1990; Sohrabi *et al.* 2011; Seegebarth *et al.* 2012; Meadows *et al.* 2013)), much of that work has involved the planner explaining its decisions with respect to its own model (i.e. current state, actions and goals) and assuming that this "*soliloquy*" also helps the human in the loop. While such a sanguine assumption may well be requited when the human is an expert "debugger" and is intimately familiar with the agent's innards, it is completely unrealistic in most human-AI interaction scenarios, where the humans may have a domain and task model that differs significantly from that used by the planner. We posit then that explanations should be seen as the robot's attempt to move the human's model to be in conformance with its own. The model reconciliation process thus forms the core of the explanation process for an XAIP agent and is thus the focus of the rest of the paper.

Our view of explanation as a model reconciliation process is supported by studies in the field of psychology which stipulate that explanations *"privilege a subset of beliefs, excluding possibilities inconsistent with those beliefs... can serve as a source of constraint in reasoning..."* (Lombrozo 2006). This is achieved in our case by the appropriate change in the expectation of the model that is believed to have engendered the plan in question. Further, authors in (Lombrozo 2012) also underline that explanations are *"typically contrastive... the contrast provides a constraint on what should figure in a selected explanation..."* - this is especially relevant in order for an explanation to be self-contained and unambiguous. Hence the requirement of optimality, which not only ensures that the current plan is valid in the updated model, but is also better than other alternatives or foils (Miller 2017).

The model reconciliation viewpoint can explain many phenomena in both explanation and transparency – e.g. the fact that well-performing, efficient teams require less, not more, explicit communication (Entin and Serfaty 1999) and the characteristics of effective team debriefing (Tannenbaum and Cerasoli 2013) after a mission or project.

The optimality criterion, and argumentation over the human mental model, makes the problem fundamentally different from model change algorithms in (Göbelbecker *et al.* 2010; Herzig *et al.* 2014; Eiter *et al.* 2010; Bryce *et al.* 2016; Porteous *et al.* 2015) which focus more on the feasibility of plans or correctness of domains.

The Model Reconciliation Process

The explanation process, in response to a plan π that the robot has come up with and is perceived as inexplicable by the human, begins with the following question –

Q: Why not a different plan $\hat{\pi}$?

This questions can arise due to one or both of two causes –

- $\mathcal{M}_h^R$, i.e. the human's approximation of the robot model is wrong. Here, since it knows its own ground truth model, the robot can use an approximation of the human mental model (known unknown) to perform model reconciliation so that both of them are on the same page.
- $\mathcal{M}_r^H$, i.e. the robot's approximation of the human model is wrong. The above approach would not work here since the robot does not know what it does not know (i.e. the real human model is an unknown unknown). However, if the above approach fails to provide a satisfactory response from the human, then the robot can conclude it must be because of this and seek out more information to update its understanding of the human model.

For the first case, the model reconciliation approach would be to provide an (1) explanation or model update $\mathcal{E}$ such that the (2) robot optimal plan is (3) feasible and at least as good as the foil in the updated model, i.e.

(1) $\widehat{\mathcal{M}}_h^R \longleftarrow \mathcal{M}_h^R + \mathcal{E}$; and

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

(3) $\delta_{\widehat{\mathcal{M}}_h^R}(\widehat{\mathcal{I}}_h^R, \pi) \models \widehat{\mathcal{G}}_h^R \ \wedge\ C(\pi, \widehat{\mathcal{M}}_h^R) < C(\hat{\pi}, \widehat{\mathcal{M}}_h^R)$.

The question can also be posed in the following form –

Q: Why plan π?

This, in essence, involves an implicit quantifier over all possible foils. Condition (3) above then must ensure that plan π is now also optimal in the updated mental model –

(3) $C(\pi, \widehat{\mathcal{M}}_h^R) = C^*_{\widehat{\mathcal{M}}_h^R}$.

In (Chakraborti *et al.* 2017b) we explore different model reconciliation processes considering four characteristics –

R1. **Completeness -** Explanations of a plan should be able to be compared and contrasted against other alternatives, so that no better solution exists. We enforce this property by requiring that in the updated human model the plan being explained is optimal – i.e. Conditions (3).

R2. **Conciseness -** Explanation should be concise so that they are easily understandable to the explainee. Larger an explanation is, the harder it is for the human to incorporate that information into her deliberative process.

R3. **Monotonicity -** This ensures that remaining model differences cannot change the completeness of an explanation, i.e. all aspects of the model that engendered the plan have been reconciled. This thus subsumes completeness and requires more detailed explanations.

R4. **Computability -** While conciseness deals with how easy it is for the explainee to understand an explanation, computability measures the ease of computing the explanation from the point of view of the planner.

A Minimally Complete Explanation (MCE) is the shortest explanation that satisfies conditions (1) and (2).

A Minimally Monotonic Explanation (MME) is the shortest explanation that is both complete and monotonic.

A Plan Patch Explanation (PPE) only includes (all the) model updates pertaining to actions in the plan π.

A Model Patch Explanation (MPE) includes all the model updates $|\mathcal{M}^R \Delta \mathcal{M}_h^R|$.

The requirements outlined above are thus often at odds - an explanation that is very easy to compute may be very hard to comprehend (c.f. Table 1). A detailed account of these explanations can be found in (Chakraborti *et al.* 2017b); we will concentrate on MCEs for the rest of the paper.

Remark Note that during model reconciliation process, the robot model need not be the ground truth. However, the robot can only explain with respect to what it believes to be true. This can, of course, be wrong and be refined iteratively through interaction with the human, as demonstrated in a decision support setting in (Sengupta *et al.* 2017).

Table 1: Requirements for different types of explanations.

Explanation Type	R1	R2	R3	R4
Plan Patch Explanation	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?

Remark We insisted that explanations must be compatible with the planners model. If this is relaxed, it allows the planner to generate "explanations" that it knows are false and deceive the human. In recent work (Chakraborti and Kambhampati 2018), we have shown that participants in a study were generally positive towards lying for the greater good especially when those actions would not be determined by their teammate, but is loath to suspend normative behavior, robot or not, in the event that they would be caught in that act (unless the robot is the recipient of the misinformation!).

Remark While in this line of work, we are concerned more with the generation of the *content* of explanations rather than the actual delivery of this information, there has been some recent work to this end. Depending on the type of interaction between the planner and the human, this can be achieved by means of natural language dialog (Perera *et al.* 2016), in the form of a graphical user interface (Sengupta *et al.* 2017; Chakraborti *et al.* 2018b) or even in mixed-reality interfaces (Chakraborti *et al.* 2018d; 2018c).

How to chose between Explicability/Explanations?

The two processes of explanations and explicability are intrinsically related in an agent's deliberative process (c.f. Figure 5) – it can generate a explicable plan to the best of its ability or it can provide explanations whenever required, or it can even opt for a combination of both if the expected human plan is too costly in its own 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). In the following discussion, we try to attain the sweet spot between plan explanations and explicability during the decision making process.

From the perspective of design of autonomy, the explicability versus explanations trade-off has two interesting implications – (1) the agent can now not only explain but also *plan* in the multi-model setting with the trade-off between compromise on its optimality and possible explanations in mind; and (2) the argumentation process is known to be a crucial function of the reasoning capabilities of humans (Mercier and Sperber 2010), and now by extension of autonomous agents as well, as a result of these algorithms that incorporate the explanation generation process into the decision making process itself. General argumentation frameworks for resolving disputes over plans have indeed been explored before (Belesiotis *et al.* 2010; Emele *et al.* 2011). Other forms of argumentation (Russell and Wefald 1991) has been aimed at meta-level reasoning of resource usage or cost of solutions. Our work can be seen as the specific case where the argumentation process is over a set of constraints that trade-off the quality of a plan and the cost of explaining it. *This is, in fact, the first of its kind algorithm that can achieve this.*

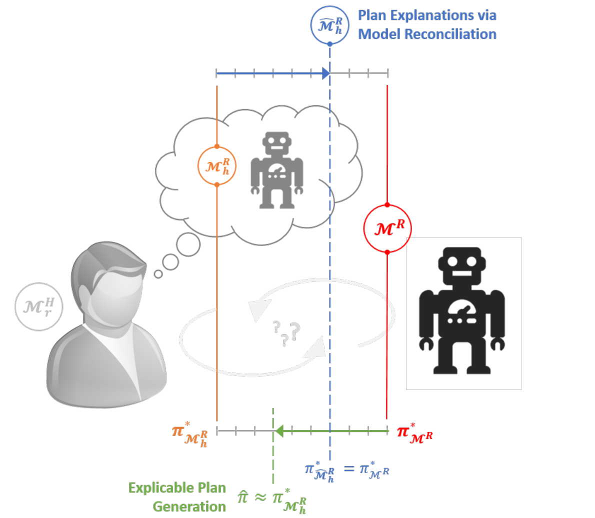

Figure 3: Balancing explicability and explanations in HAP.

The result of a trade off in the relative cost of explicability and explanations during the plan generation process 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 to the human, is traded off according to a constant α –

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

(2) $\widehat{\mathcal{M}}_h^R \longleftarrow \mathcal{M}_h^R + \mathcal{E}$;

(3) $C(\pi, \widehat{\mathcal{M}}_h^R) = C^*_{\widehat{\mathcal{M}}_h^R}$; and

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

With higher values of α the planner generates plans that require more explanation; with lower α it will generate more explicable plans. Thus, using this hyperparameter, an autonomous agent can deliberate over the trade-off in the costs it incurs in being explicable to the human (second minimizing term in (4)) versus explaining its decisions (first minimizing term in (4)). Note that this trade-off is irrespective of the cognitive burden of those decisions on the human in the loop. For example, for a robot in a collapsed building during a search and rescue task, may have limited bandwidth for communication and hence prefer to be explicable instead.

Demonstration Figure 2:A illustrates a section of the environment where this whole scenario plays out. The orange marks indicate rubble that has blocked a passage, while the green marks indicate collapsed walls. The robot, currently located at the position marked with a blue **O**, is tasked with taking a picture at location marked with an orange **O** in the figure. The external commander's expects the robot to take the path shown in red, which is no longer possible. The robot armed with `MEGA*` 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. A video demonstration can be viewed at https://youtu.be/Yzp4FU6Vn0M. The first part of the video demonstrates the plan generated by MEGA* for low α values. As expected, it chooses the blue path that requires the least amount of explanation, i.e. the most explicable plan. In fact, the robot only needs to explain a single initial state change to make its plan optimal –

```
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. 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 plan (shown in green). Indeed, MEGA* switches to the robot's optimal plan for higher values of α along with the following explanation –

```
add-has-initial-state-clear_path p6 p7
add-has-initial-state-clear_path p7 p5
remove-has-initial-state-clear_path p1 p8
```

What happens if the mental model is unknown?

The model reconciliation process described above is only feasible if inconsistencies of the robot model with the human mental model are known precisely. Although we made this assumption so far as a first step towards formalizing the model reconciliation process, this can be hard to achieve in practice. Instead, the agent may end up having to explain its decisions with respect to a *set of possible models* which is its best estimation of the human's knowledge state learned in the process of interactions. In this situation, the robot can try to compute MCEs for each possible configuration. However, this can result in situations where the explanations computed for individual models independently are not consistent across all possible target domains. Thus, in the case of model uncertainty, such an approach cannot guarantee that the resulting explanation will be acceptable. Instead, we want to find an explanation such that $\forall i\ \pi^*_{\widehat{\mathcal{M}}^R_{h_i}} \equiv \pi^*_{\mathcal{M}^R}$.

This is a single model update that makes the given plan optimal (and hence explained) in all the updated domains (or in all possible domains). At first glance, it appears that such an approach, even though desirable, might turn out to be prohibitively expensive especially since solving for a *single* MCE involves search in the model space where each search node is an optimal planning problem (Chakraborti *et al.* 2017b). However, it turns out that the same search strategy can be employed here as well by representing the human mental model as an *annotated* model (Sreedharan *et al.* 2018a). Condition (3) for an MCE now becomes –

(3) $C(\pi, g(\mathbb{M}^R_h)) = C^*_{g(\mathbb{M}^R_h)}$

This is hard to achieve since it is not known which is the actual mental model of the human. So we want to preserve the optimality criterion for all (or as many) instantiations of the incomplete estimation of the explainee's mental model. Keeping this in mind, we define *robustness* of an explanation for an incomplete mental models as the probability mass of models where it is a valid explanation.

Robustness of an explanation $\mathcal{E}$ is given by –

$$R(\mathcal{E}) = \sum_{inst(\widehat{\mathcal{M}}^R_h) \text{ s.t. } C(\pi, inst(\widehat{\mathcal{M}}^R_h)) = C^*_{inst(\widehat{\mathcal{M}}^R_h)}} \mathcal{L}(inst(\widehat{\mathcal{M}}^R_h))$$

A Conformant Explanation is such that $R(\mathcal{E}) = 1$.

This means a conformant explanation ensures that the given plan is explained in all the models in the completion set of the human model.

Demonstration. Consider now that the robot is located at P1 (blue) and needs to collect data from P5 (c.f. Figure 2:C). While the human commander understands the goal, she is under the false impression that the paths from P1 to P9 and P4 to P5 are unusable (red question marks). She is also unaware of the robot's inability to use its hands. On the other hand, while the robot does not have a complete picture of the human's mental model, it understands that any differences between the models are related to (1) Path from P1 to P9; (2) Path from P4 to P5; (3) Robot's ability to use its hands; and (4) Whether the Robot needs its arm to clear rubble. Thus, from the robot's perspective, the human model can be one of sixteen possible models (one of which is the actual mental model). Here, a conformant explanation for the optimal robot plan (blue) is as follows (a demonstration can be viewed at https://youtu.be/bLqrtffW6Ng) –

```
remove-known-INIT-has-add-effect-hand_capable
add-annot-clear_passage-has-precondition-hand_capable
remove-annot-INIT-has-add-effect-clear_path P1 P9
```

Note that conformant explanations can contain superfluous information – i.e. asking the human to remove nonexistent conditions or add existing ones. In the previous example, the second explanation (regarding the need of the hand to clear rubble) was already known to the human and was thus superfluous information. Such redundant information can be annoying and may end up reducing the human's trust in the robot. This can be avoided by –

- Increasing the cost of model updates involving uncertain conditions relative to those involving known preconditions or effects. This ensures that the search prefers explanations that contain known conditions. By definition, such explanations will not have superfluous information.
- However, such explanations may not exist. Instead, we can convert conformant explanations into *conditional* ones by turning each model update for an annotated condition into a question and only provide an explanation if the human's response warrants it – e.g. instead of asking the human to update the precondition of `clear_passage`, the robot can first ask if the human thinks that action has a precondition `hand_usable`.

Thus, one way of removing superfluous explanations is to engage the human in conversation and reduce the size of the completion set. Consider the following exchange –

```
R : Are you aware that the path from P1 to P4 has collapsed?
H : Yes.
> R realizes the plan is optimal in all possible models.
> It does not need to explain further.
```

A Conditional Explanation is represented by a policy that maps the annotated model (represented by a $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$ model pair) to either a question regarding the existence of a condition in the human ground model or a model update request. The resultant annotated model is produced, by either applying the model update directly into the current model or by updating the model to conform to human's answer regarding the existence of the condition.

Note that in asking questions such as these, the robot is trying to exploit the human's (lack of) knowledge of the problem in order to provide more concise explanations. This can be construed as a case of lying by omission and can raise interesting ethical considerations (Chakraborti and Kambhampati 2018). Humans, during an explanation process, tend to undergo this same "selection" process (Miller 2017) as well in determining which of the many reasons that could explain an event is worth highlighting. It is worthwhile investigating similar behavior for autonomous agents.

Anytime Explanations Since dealing with model uncertainty can be computationally expensive, we relax the minimality requirement and introduce an anytime depth first explanation generation algorithm. This is explained in detail in (Sreedharan *et al.* 2018a).

What if there are multiple humans in the loop?

While generating explanations for a *set of models*, the robot is essentially trying to cater to multiple human models at the same time. We posit then that the same approaches can be adopted to situations when there are *multiple humans* in the loop instead of a single human whose model is not known with certainty. As before, computing separate explanations (Chakraborti *et al.* 2017b) for each agent can result in situations where the explanations computed for individual models independently are not consistent across all the possible target domains. In the case of multiple teammates being explained to, this may cause confusion and loss of trust, especially in teaming with humans who are known (Cooke *et al.* 2013) to rely on shared mental models. Thus *conformant explanations* can find useful applications in dealing with not only model uncertainty but also model multiplicity.

In order to do this, from the set of target human mental models we construct an annotated model so that *the preconditions and effects that appear in all target models become necessary ones, and those that appear in just a subset are possible ones*. As before, we find a single explanation that is a satisfactory explanation for the entire set of models, without having to repeat the standard MRP process over all possible models while coming up with an explanation that can satisfy all of them and thus establish common ground (Chakraborti *et al.* 2018c; Sreedharan *et al.* 2018b).

Demonstration We go back to our use case, now with *two* human teammates, one external and one internal. A video of the demonstration is available at https://youtu.be/hlPTmggRTQA. The robot is now positioned at `P1` and is expected to collect data from location `P5`. Before the robot can perform its `surveil` action, it needs to obtain a set of tools from the internal human agent. The human agent is initially located at `P10` and is capable of traveling to reachable locations to meet the robot for the handover. Here the external commander incorrectly believes that the path from `P1` to `P9` is clear and while the one from `P2` to `P3` is closed. The internal human agent, on the other hand, not only believes in the errors mentioned above but is also under the assumption that the path from `P4` to `P5` is not traversable. Due to these different initial states, each of these agents ends up generating a different optimal plan. The plan expected by the external commander requires the robot to move to location `P10` (via `P9`) to meet the human. After collecting the package from the internal agent, the commander expects it to set off to `P5` via `P4`. The internal agent, on the other hand, believes that he needs to travel to `P9` to hand over the package. As he believes that the corridor from `P4` to `P5` is blocked, he expects the robot to take the longer route to `P5` through `P6`, `P7`, and `P8` (orange). Finally, the optimal plan for the robot (blue) involves the robot meeting the human at `P4` on its way to `P5`. Using `MEGA*-Conformant`, we find the smallest explanation, which can explain this plan to both humans –

```
add-INIT-has-clear_path P4 P5
remove-INIT-has-clear_path P1 P9
add-INIT-has-clear_path P2 P3
```

While the last two model changes are equally relevant for both the agents, the first change is specifically designed to help the internal. The first update helps convince the human that the robot can indeed reach the goal through `P4`, while the next two help convince both agents as to why the agents should meet at `P4` rather than other locations.

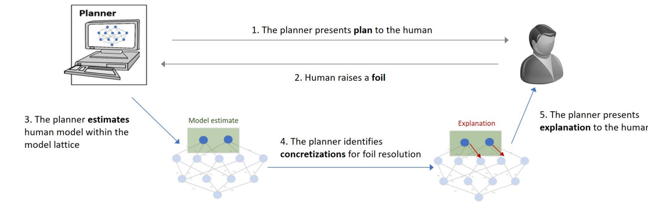

Figure 4: Hierarchical explanation generation approach.

How to model human expertise?

Most of the above discussion has focused on generating explanations in cases where both the human and the robot understands the task at the same granularity. Applying model reconciliation without acknowledging the difference in the level of "expertise" can lead to confusion and information overload. Indeed, explanation generation techniques for machine learning systems have explicitly modeled this difference (Ribeiro *et al.* 2016; 2018).

In (Sreedharan *et al.* 2018c), authors have looked at ways of generating explanations when the human has access to only an abstract version of the model of the robot. Specifically, they focus on state abstractions where the abstract model was formed by projecting out a certain subset of state fluents (Srivastava *et al.* 2016), though the principles are likely to carry over to other types of abstraction as well (e.g. temporal abstractions of the types discussed in (Marthi *et al.* 2007)). Since the abstract model may be logically weaker, the human may incorrectly believe that an optimal plan suggested by the planner is suboptimal. When presented with the plan π, the human can respond by either presenting a set of *foils* F. In such cases, the *explanation* takes the form of information about a set of state properties which when introduced into the human model resolves or invalidates the set of foils. Thus, the explanation can be uniquely represented by a sequence of propositions $\epsilon = \langle p_1, .., p_k \rangle$ as follows –

(1) A set of foils $F = \{\pi_1, .., \pi_m\}$ such that $\forall \pi \in F$, $\delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \not\models \mathcal{G}^R$ and $\delta_{\mathcal{M}_h^R}(\mathcal{I}_h^R, \pi) \not\models \mathcal{G}_h^R$
(2) An explanation $\epsilon = \langle p_1, ..., p_k \rangle$, such that $\widehat{\mathcal{M}}_h^R \longleftarrow \mathcal{M}_h^R + \mathcal{E}$;
(3) $\forall \pi \in F$, $\delta_{\widehat{\mathcal{M}}_h^R}(\mathcal{I}_h^R, \pi) \not\models \mathcal{G}_h^R$

One of the main challenges with this method is the uncertainty about the human model. To address this, the authors build a lattice of possible models from the task model called *model lattice* ($\mathcal{L}$) as shown in Figure 4. The lattice consists of possible abstractions of the concrete task model and an edge exists between two models if there exists a single predicate that can be projected from one model to create the other. The foils are used to estimate the possible human model and use this estimate to compute the explanations.

How do humans reconcile models?

The design of "human-aware" algorithms is, of course, incomplete without evaluations of the same with actual humans in the loop. Thus, in the final part of this discussion, we will report on the the salient findings from a controlled user study we undertook recently in order to evaluate the usefulness of the model reconciliation approach. A detailed report of the study can be read at (Chakraborti *et al.* 2018a).

Experimental Setup For the study, we exposed the external commanders interface (c.f. Figure 2:D) to participants who, based on their map (which they are told may differ from the robots) had to identify if a given plan (which may be optimal in the robot model or explicable or even balanced) looks optimal or satisficing to them. If the player is unsure, they can ask for an explanation. The explanations provided are one of the types described before.

Study-1 In the first set of experiments, participants assumed the role of the explainer. It was found that, when left to themselves, they generated explanations of the type MPE or (if communication was restricted) MCE. Further, in subjective responses, they considered model reconciliation as necessary and sufficient for the explanation process.

Study-2 Here, participants assumed the role of the explainer, and had to identify, on the basis of explanations provided the quality of the given plan. We found that the participants were indeed able to distinguish between optimal plans (when provided with MCEs or MPEs) and (perceived) satisficing plans (when provided with PPEs) and were in general overwhelmingly in favor of model reconciliation as an effective form of explanation. We further found that explicable plans indeed reduced the call of explanations, while balanced plans preserved their outlook towards the explanations while allowing the robot to trade-off its communication cost with the optimality of its plans.

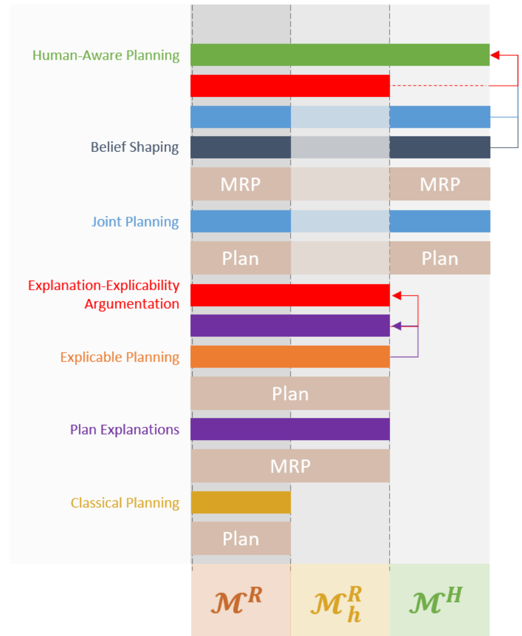

Figure 5: A subsumption architecture for HAP.

Work in Progress

So far we have have not considered differences in the cost function which, though falls under the scope of model differences to be explained, can introduce interesting challenges to the model reconciliation problem. It may be possible to learn such functions through interactions with the human as in (Zhang *et al.* 2017; Kulkarni *et al.* 2016). Further, we have not modeled the computational capability of the human which also affect the process – this can be potentially handled by modeling $\epsilon-$optimal humans or by considering top-K plans (Katz *et al.* 2018) while checking the optimality condition during model space search. Finally, we do not consider system level constraints (such as time and resources) in the explanations which remain out of scope of explanations viewed as a model reconciliation process.

Conclusion

The different behaviors engendered by multi-model argumentation can be *composed* to form more and more sophisticated forms of human-aware behavior. We thus conclude with a hierarchical composition of behaviors in the form of a subsumption architecture for human-aware planning, inspired by (Brooks 1986). This is illustrated in Figure 5. The basic reasoning engines are the Plan and MRP (Model Reconciliation) modules. The former accepts model(s) of planning problems and produces a plan, the latter accepts the same and an produces a new model. The former operates in plan space and gives rise to classical, joint and explicable planning depending on the models it is operating on, while the latter operates in model space to produce explanations and belief shaping behavior. These are then composed to form argumentation modules for trading of explanations and explicability and human-aware planning in general.

Acknowledgements This research is supported in part by the AFOSR grant FA9550-18-1-0067, the ONR grants N00014-16-1-2892, N00014-13-1-0176, N00014-13-1-0519, N00014-15-1-2027, N00014-18-1-2442 and the NASA grant NNX17AD06G. The first author is also supported by the IBM Ph.D. Fellowship from 2016-18.

References

Rachid Alami, Aurélie Clodic, Vincent Montreuil, Emrah Akin Sisbot, and Raja Chatila. Toward Human-Aware Robot Task Planning. In *AAAI Spring Symposium: To Boldly Go Where No Human-Robot Team Has Gone Before*, 2006.

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.

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

Alexandros Belesiotis, Michael Rovatsos, and Iyad Rahwan. Agreeing on plans through iterated disputes. In *AAMAS*, pages 765–772, 2010.

Rodney Brooks. A robust layered control system for a mobile robot. *IEEE journal on robotics and automation*, 2(1):14–23, 1986.

Dan Bryce, J. Benton, and Michael W. Boldt. Maintaining Evolving Domain Models. In *IJCAI*, 2016.

T. Chakraborti and S. Kambhampati. Algorithms for the Greater Good! On Mental Modeling and Acceptable Symbiosis in Human-AI Collaboration. *ArXiv e-prints*, January 2018.

T. Chakraborti, G. Briggs, K. Talamadupula, Yu Zhang, M. Scheutz, D. Smith, and S. Kambhampati. Planning for serendipity. In *IROS*, 2015.

Tathagata Chakraborti, Yu Zhang, David Smith, and Subbarao Kambhampati. Planning with Resource Conflicts in Human-Robot Cohabitation. In *AAMAS*, 2016.

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

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

T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati. Plan Explanations as Model Reconciliation – An Empirical Study. *ArXiv e-prints*, February 2018.

Tathagata Chakraborti, Fadnis P. Kshitij, Kartik Talamadupula, Mishal Dholakia, Biplav Srivastava, Jeffrey O Kephart, and Rachel KE Bellamy. Visualizations for an explainable planning agent. *ICAPS UISP*, 2018.

Tathagata Chakraborti, Sailik Sengupta, and Subbarao Kambhampati. MA-RADAR – A Mixed-Reality Interface for Collaborative Decision Making. *ICAPS UISP*, 2018.

Tathagata Chakraborti, Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace. In *HRI 2018 Workshop on Virtual, Augmented and Mixed-Reality for Human-Robot Interactions (VAM-HRI'18)*, 2018.

Marcello Cirillo, Lars Karlsson, and Alessandro Saffiotti. Human-aware Task Planning: An Application to Mobile Robots. *ACM Transactions on Intelligent Systems and Technology*, 2010.

Marcello Cirillo. *Planning in inhabited environments: human-aware task planning and activity recognition*. PhD thesis, Örebro university, 2010.

Nancy J Cooke, Jamie C Gorman, Christopher W Myers, and Jasmine L Duran. Interactive team cognition. *Cognitive science*, 37(2):255–285, 2013.

Thomas Eiter, Esra Erdem, Michael Fink, and Ján Senko. Updating action domain descriptions. *Artificial intelligence*, 2010.

Chukwuemeka D Emele, Timothy J Norman, and Simon Parsons. Argumentation strategies for plan resourcing. In *AAMAS*, 2011.

Elliot E Entin and Daniel Serfaty. Adaptive team coordination. *Human factors*, 41(2):312–325, 1999.

Maria Fox, Derek Long, and Daniele Magazzeni. Explainable Planning. In *IJCAI XAI Workshop*, 2017.

Moritz Göbelbecker, Thomas Keller, Patrick Eyerich, Michael Brenner, and Bernhard Nebel. Coming up with good excuses: What to do when no plan can be found. 2010.

Andreas Herzig, Viviane Menezes, Leliane Nunes de Barros, and Renata Wassermann. On the revision of planning tasks. In *ECAI*, 2014.

Subbarao Kambhampati. A classification of plan modification strategies based on coverage and information requirements. In *AAAI 1990 Spring Symposium on Case Based Reasoning*, 1990.

Michael Katz, Shirin Sohrabi, Octavian Udrea, and Dominik Winterer. A Novel Iterative Approach to Top-k Planning. In *ICAPS*, 2018.

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

Anagha Kulkarni, Tathagata Chakraborti, Yantian Zha, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. Explicable Robot Planning as Minimizing Distance from Expected Behavior. *CoRR*, abs/1611.05497, 2016.

Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. Implicit robot-human communication in adversarial and collaborative environments. *CoRR*, abs/1802.06137, 2018.

Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable Agency for Intelligent Autonomous Systems. In *AAAI/IAAI*, 2017.

Tania Lombrozo. The Structure and Function of Explanations . *Trends in Cognitive Sciences*, 10(10):464 – 470, 2006.

Tania Lombrozo. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276, 2012.

Bhaskara Marthi, Stuart J Russell, and Jason Andrew Wolfe. Angelic semantics for high-level actions. In *ICAPS*, pages 232–239, 2007.

Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. PDDL-the planning domain definition language. 1998.

Ben Leon Meadows, Pat Langley, and Miranda Jane Emery. Seeing beyond shadows: Incremental abductive reasoning for plan understanding. In *AAAI Workshop: Plan, Activity, and Intent Recognition*, volume 13, page 13, 2013.

Hugo Mercier and Dan Sperber. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 2010.

Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *CoRR*, abs/1706.07269, 2017.

Tuan Anh Nguyen, Minh Do, Alfonso Emilio Gerevini, Ivan Serina, Biplav Srivastava, and Subbarao Kambhampati. Generating diverse plans to handle unknown and partially known user preferences. *Artificial Intelligence*, 2012.

Tuan Nguyen, Sreedharan, and Subbarao Kambhampati. Robust planning with incomplete domain models. *Artificial Intelligence*, 245:134 – 161, 2017.

Vittorio Perera, Sai P. Selvaraj, Stephanie Rosenthal, and Manuela Veloso. Dynamic Generation and Refinement of Robot Verbalization. In *RO-MAN*, Columbia University, NY, August 2016.

Julie Porteous, Alan Lindsay, Jonathon Read, Mark Truran, and Marc Cavazza. Automated extension of narrative planning domains with antonymic operators. In *AAMAS*, pages 1547–1555, 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144. ACM, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018.

Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, 2003.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial intelligence*, 49(1-3):361–395, 1991.

Bastian Seegebarth, Felix Müller, Bernd Schattenberg, and Susanne Biundo. Making hybrid plans more clear to human users-a formal approach for generating sound explanations. In *ICAPS*, 2012.

Sailik Sengupta, Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. Radar-a proactive decision support system for human-in-the-loop planning. In *AAAI Fall Symposium on Human-Agent Groups*, 2017.

Shirin Sohrabi, Jorge A. Baier, and Sheila A. McIlraith. Preferred explanations: Theory and generation via planning. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI-11)*, pages 261–267, San Francisco, USA, August 2011.

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

Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Handling model uncertainty and multiplicity in explanations via model reconciliation. 2018.

Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. Hierarchical expertise-level modeling for user specific robot-behavior explanations. *arXiv preprint arXiv:1802.06895*, 2018.

Biplav Srivastava, Tuan Anh Nguyen, Alfonso Gerevini, Subbarao Kambhampati, Minh Binh Do, and Ivan Serina. Domain independent approaches for finding diverse plans. In *IJCAI*, pages 2016–2022, 2007.

Siddharth Srivastava, Stuart J Russell, and Alessandro Pinto. Metaphysics of planning domain descriptions. In *AAAI*, 2016.

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.

Scott I Tannenbaum and Christopher P Cerasoli. Do team and individual debriefs enhance performance? a meta-analysis. *Human factors*, 55(1):231–245, 2013.

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.

Daniel S Weld and Gagan Bansal. Intelligible artificial intelligence. *arXiv preprint arXiv:1803.04263*, 2018.

Yu Zhang, Vignesh Narayanan, Tathagata Chakraborty, and Subbarao Kambhampati. A human factors analysis of proactive assistance in human-robot teaming. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

Yu Zhang, Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan Explicability for Robot Task Planning. In *RSS Workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics*, 2016.

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