

Plan Explanations as Model Reconciliation: Moving Beyond *Explanation as Soliloquy*

Tathagata Chakraborti* and Sarath Sreedharan* and Yu Zhang and Subbarao Kambhampati

School of Computing, Informatics, and Decision Systems Engineering

Arizona State University, Tempe, AZ 85281 USA

{ tchakra2, ssreedh3, yzhan442, rao } @ asu.edu

Abstract

When AI systems interact with humans in the loop, they are often called on to provide explanations for their plans and behavior. Past work on plan explanations primarily involved the AI system explaining the correctness of its plan and the rationale for its decision in terms of its own model. Such soliloquy is wholly inadequate in most realistic scenarios where the humans have domain and task models that differ significantly from that used by the AI system. We posit that the explanations are best studied in light of these differing models. In particular, we show how explanation can be seen as a “model reconciliation problem” (MRP), where the AI system in effect suggests changes to the human’s model, so as to make its plan be optimal with respect to that changed human model. We will study the properties of such explanations, present algorithms for automatically computing them, and evaluate the performance of the algorithms.

1 Introduction

There has been significant renewed interest recently in developing AI systems that can automatically provide explanations to humans in the loop. While much of the interest has been focused on learning systems that can explain their classification decisions, a related broader problem involves providing explanations in the context of human-AI interaction and human-in-the-loop decision making systems. In such scenarios, the automated agents are called upon to provide explanation of their behavior or plans [Langley, 2016].

Although explanation of plans has been investigated in the past (c.f. [Kambhampati, 1990; Sohrabi *et al.*, 2011]), much of that work 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 required 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 AI system. This is illustrated in Figure 1,

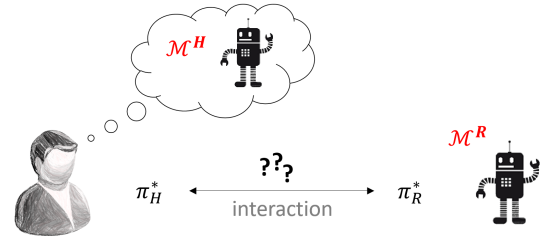


Figure 1: Interaction between humans and AI systems is best analyzed in light of their differing models. The robot here generates an optimal plan π_R^* with respect to its model \mathcal{M}^R , which is interpreted by the human with respect to his model \mathcal{M}^H . Explanations are needed when π_R^* is *not* an optimal plan with respect to \mathcal{M}^H .

where the plans generated by the AI system with respect to its model need to be interpreted by the human with respect to his model. Of course, the AI system can avoid the need to provide explanations by being “*explicable*” [Zhang *et al.*, 2017; Kulkarni *et al.*, 2016] - i.e., generate plans that also make sense with respect to the humans’ model. Such explicability requirement however puts additional constraints on the agent’s plans, and may not always be feasible. When the robot’s plan is different from what the human would expect given his model of the world, the robot will be called on to “explain” its plan. We posit that such explanations should be seen as the robot’s attempt to move the human’s model to be in conformance with its own.

The primary contribution of this paper is to show how such model updates or *explanations* can be formulated concisely as the *model reconciliation problem* (MRP), which aims to make minimal changes to the human’s model to bring it closer to the robot’s model, in order to make the robot’s plan optimal with respect to this changed human’s model. One immediate complication in tackling an MRP is that the human’s model is not directly made available to the robot, and will have to be learned instead (c.f. [Zhang *et al.*, 2017]). The learned model may also be in a different form and at a different level of abstraction than the one used by the robot [Tian *et al.*, 2016; Perera *et al.*, 2016]. Nevertheless, for the purposes of this paper, we will assume that the human’s model is made available and is in PDDL format, just like the robot’s one. This allows us to focus on the explanation generation aspects.

In the rest of the paper, we will formalize the scenario in Figure 1 as the *Multi-Model Planning* setting, and character-

* Authors marked with asterisk contributed equally.



Figure 2: The Fetch in the crouched position with arm tucked (left), torso raised and arm outstretched (middle) and the rather tragic consequences of a mistaken action model (right).

ize explanation generation as a *model reconciliation process* in it. We start by enumerating a few desirable requirements of such explanations - namely completeness, conciseness, monotonicity and computability. We then formulate different kinds of explanations that satisfy these requirements and relax one of these requirements at a time in the interests of computability. We present an A*-search formulation for searching in the space of models to compute these explanations, and develop approximations and heuristics for the same. Finally, we present a preliminary evaluation of the efficiency of our algorithms for generating explanations in randomly generated problems in a few benchmark planning domains.

A Motivating Example Let us illustrate the concept of explanations via model reconciliation through an example based on the Fetch robot whose design requires it to *tuck* its arms and lower its torso or *crouch* before moving - which is not obvious to a human navigating it. This may lead to an unbalanced base and toppling of the robot if the human deems such actions as unnecessary. The move action for the robot is described in PDDL in the following model snippet -

```
(:action move
:parameters      (?from ?to - location)
:precondition     (and (robot-at ?from)
                      (hand-tucked) (crouched))
:effect           (and (robot-at ?to)
                      (not (robot-at ?from))))

(:action tuck
:parameters      ()
:precondition     ()
:effect           (and (hand-tucked)
                      (crouched)))

(:action crouch
:parameters      ()
:precondition     ()
:effect           (and (crouched)))
```

Notice that the *tuck* action also involves a lowering of torso so that the arm can rest on the base once it is tucked in. Now, consider a problem with the following initial and goal states (here, identical for both the robot and the human) -

```
(:init (block-at b1 loc1) (robot-at loc1) (hand-empty))
(:goal (and (block-at b1 loc2)))
```

An optimal plan for the robot, in this case, involves a *tuck* action followed by a *move* -

```
pick-up b1 -> tuck -> move loc1 loc2 -> put-down b1
```

The human, on the other hand, expects a much simpler model, as shown below. The *move* action does not have the preconditions for tucking the arm and lowering the torso, while *tuck* does not automatically lower the torso either.

```
(:action move
:parameters      (?from ?to - location)
:precondition     (and (robot-at ?from)
                      (not (robot-at ?to)))
:effect           (and (robot-at ?to)
                      (not (robot-at ?from))))

(:action tuck
:parameters      ()
:precondition     ()
:effect           (and (hand-tucked)))

(:action crouch
:parameters      ()
:precondition     ()
:effect           (and (crouched)))
```

Clearly, the original plan is no longer optimal (and hence explicable) here. One possible model update (i.e. explanation) that can mitigate this situation is -

```
Explanation >> MOVE_LOC1_LOC2-has-precondition-HAND-TUCKED
```

This correction brings the human and the robot model closer, and is necessary and sufficient to make the robot's plan optimal in the resultant domain. As indicated before, we refer to such model corrections as *multi-model explanations*.

2 Related Work

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 in our explanations, which not only ensures that the current plan is valid in the updated model, but is also better than other alternatives. This is consistent with the notion of optimal (single-model) explanations investigated in [Sohrabi *et al.*, 2011] where less costly plans are referred to as *preferred explanations*. The optimality criterion, however, makes the problem fundamentally different from model change algorithms in [Göbelbecker *et al.*, 2010; Herzig *et al.*, 2014; Bryce *et al.*, 2016] which focus more on the feasibility of plans or correctness of domains.

Finally, while the human-in-the-loop setting discussed here does bring back memories of mixed-initiative planners of the past [Ferguson *et al.*, 1996; Ai-Chang *et al.*, 2004], most of the work there involved the humans entering the land of planners; and not the other way around. Not surprisingly, it did not have the planner taking the human model into account in its planning or explanation.

3 The Multi-Model Planning (MMP) Setting

A Classical Planning Problem [Russell *et al.*, 2003] requires a model $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}, \mathcal{G} \rangle$ (represented in PDDL [McDermott *et al.*, 1998]) consisting of the domain $\mathcal{D} = \langle F, A \rangle$ - where F is a finite set of fluents that define the world state $s \subseteq F$, and A is a finite set of actions - and the initial and goal states $\mathcal{I}, \mathcal{G} \subseteq F$. Action $a \in A$ is a tuple $\langle c_a, pre(a), eff^+(a), eff^-(a) \rangle$ where c_a

denotes cost, and $pre(a), eff^+(a), eff^-(a) \subseteq F$ is the set of preconditions and add / delete effects, i.e. $\delta_{\mathcal{M}}(s, a) \models \perp$ if $s \not\models pre(a)$; else $\delta_{\mathcal{M}}(s, a) \models s \cup eff^+(a) \setminus eff^-(a)$ where $\delta_{\mathcal{M}}(\cdot)$ is the transition function. The cumulative transition function is given by $\delta_{\mathcal{M}}(s, \langle a_1, a_2, \dots, a_n \rangle) = \delta_{\mathcal{M}}(\delta_{\mathcal{M}}(s, a_1), \langle a_2, \dots, a_n \rangle)$. The solution to the planning problem is a sequence of actions or a (satisficing) plan $\pi = \langle a_1, a_2, \dots, a_n \rangle$ 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 $\pi^* = \arg \min_{\pi} \{C(\pi, \mathcal{M}) \mid \forall \pi \text{ such that } \delta_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}\}$ is called the (cost) optimal plan. We will refer to the cost of the optimal plan in the model \mathcal{M} as $C_{\mathcal{M}}^*$.

From the perspective of classical planning, the question of **plan explicability** may indeed be posed in terms of plan optimality. The intuition here is that if a plan is the best possible plan that the planner could have come up with, then it is also comprehensible to a human *with the same planning model and reasoning capabilities*, i.e. the existence of such a plan requires no further explanation. However, the human's understanding of a planning problem often differs from the planner's, e.g. if she does not have access to the planner's actual goal, or does not know the current state accurately, or has a different perception of the action model being used. In such situations, a plan π produced in the robot's model \mathcal{M}^R is being evaluated in terms of a different model \mathcal{M}^H in the human's mind, as a result of which what is optimal (and explicable) in the planner's model may no longer be so in the human's, i.e. $C(\pi, \mathcal{M}^R) = C_{\mathcal{M}^R}^*$, but $C(\pi, \mathcal{M}^H) > C_{\mathcal{M}^H}^*$. Based on this, we define the following setting -

A Multi-Model Planning (MMP) Setting is given by the tuple $\langle \mathcal{M}^R, \mathcal{M}^H \rangle$, where $\mathcal{M}^R = \langle D^R, \mathcal{I}^R, \mathcal{G}^R \rangle$ is the planner's model of the planning problem, while $\mathcal{M}^H = \langle D^H, \mathcal{I}^H, \mathcal{G}^H \rangle$ is the human's approximation of the same.

As we mentioned in the introduction, from the point of view of the planner, there can be two approaches to achieve common ground with the human in such settings - (1) *Change its own behavior in order to be explicable to the human* - in [Zhang et al., 2017; Kulkarni et al., 2016] the authors propose to modify the robot plan π itself so that $C(\pi, \mathcal{M}^H) \approx C_{\mathcal{M}^H}^* \wedge \delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \models \mathcal{G}^R$. Thus the planner chooses to sacrifice optimality in order to make its behavior explicable to the human observer; and (2) *Bring the human's model closer to its own by means of explanations in the form of model updates* - here, the planner does not change its own behavior, but rather corrects the human's incorrect perception of its model via *explanations*. We refer to this as *the model reconciliation process*. In this paper, we will focus on this only.

The Model Reconciliation Problem (MRP) is a tuple $\langle \pi^*, \langle \mathcal{M}^R, \mathcal{M}^H \rangle \rangle$ where $C(\pi^*, \mathcal{M}^R) = C_{\mathcal{M}^R}^*$, i.e. the robot's model, the human's approximation of it, and a plan that is optimal in the former.

Before we go into the details of the model reconciliation process, we will define the following state representation over planning problems. We intend to use this in our "model-space

search" for model reconciliation.

$$\begin{aligned} \mathcal{F} = & \{init-has-f \mid \forall f \in F^H \cup F^R\} \cup \{goal-has-f \mid \forall f \in F^H \cup F^R\} \\ & \bigcup_{a \in A^H \cup A^R} \{a-has-precondition-f, a-has-add-effect-f, \\ & a-has-del-effect-f \mid \forall f \in F^H \cup F^R\} \\ & \cup \{a-has-cost-c_a \mid a \in A^H\} \cup \{a-has-cost-c_a \mid a \in A^R\}. \end{aligned}$$

A mapping function $\Gamma : \mathcal{M} \mapsto s$ represents any planning problem $\mathcal{M} = \langle \langle F, A \rangle, \mathcal{I}, \mathcal{G} \rangle$ as a state $s \subseteq \mathcal{F}$ as follows -

$$\tau(f) = \begin{cases} init-has-f & \text{if } f \in \mathcal{I}, \\ goal-has-f & \text{if } f \in \mathcal{G}, \\ a-has-precondition-f & \text{if } f \in pre(a), a \in A \\ a-has-add-effect-f & \text{if } f \in eff^+(a), a \in A \\ a-has-del-effect-f & \text{if } f \in eff^-(a), a \in A \\ a-has-cost-f & \text{if } f = c_a, a \in A \end{cases}$$

$$\begin{aligned} \Gamma(\mathcal{M}) = & \{\tau(f) \mid \forall f \in \mathcal{I} \cup \mathcal{G} \cup \\ & \bigcup_{a \in A} \{f' \mid \forall f' \in \{c_a\} \cup pre(a) \cup eff^+(a) \cup eff^-(a)\}\} \end{aligned}$$

We can now define a *model-space search problem* $\langle \langle \mathcal{F}, \Lambda \rangle, \Gamma(\mathcal{M}_1), \Gamma(\mathcal{M}_2) \rangle$ with a new action set Λ containing unit model change actions $\lambda : \mathcal{F} \rightarrow \mathcal{F}$ such that $|s_1 \Delta s_2| = 1$, where the new transition or edit function is given by $\delta_{\mathcal{M}_1, \mathcal{M}_2}(s_1, \lambda) = s_2$ such that condition 1 : $s_2 \setminus s_1 \subseteq \Gamma(\mathcal{M}_2)$ and condition 2 : $s_1 \setminus s_2 \not\subseteq \Gamma(\mathcal{M}_2)$ are satisfied. This means that model change actions can only make a single change to a domain at a time, and *all these changes are consistent with the model of the planner*. The solution to a model-space search problem is given by a set of edit functions $\{\lambda_i\}$ that can transform the model \mathcal{M}_1 to the model \mathcal{M}_2 , i.e. $\delta_{\mathcal{M}_1, \mathcal{M}_2}(\Gamma(\mathcal{M}_1), \{\lambda_i\}) = \Gamma(\mathcal{M}_2)$.

A Multi-Model Explanation denoted by \mathcal{E} is a solution to an MRP, i.e. a solution to the model-space search problem $\langle \langle \mathcal{F}, \Lambda \rangle, \Gamma(\mathcal{M}^H), \Gamma(\widehat{\mathcal{M}}) \rangle$ with the transition function $\delta_{\mathcal{M}^H, \mathcal{M}^R}$ such that $C(\pi^*, \widehat{\mathcal{M}}) - C_{\widehat{\mathcal{M}}}^* < C(\pi^*, \mathcal{M}^H) - C_{\mathcal{M}^H}^*$.

This means that in the updated model after the explanation, the plan in question is closer to the optimal (and hence less inexplicable) than it was in the original model that the human had. The human in the loop can either chose to use this explanation to update her own model or negotiate in course of further dialog (e.g. to update the robot's model).

As we go on to develop approaches to compute different types of such explanations, we will consider the following four requirements that characterize each solution.

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.

- An explanation is complete iff $C(\pi^*, \widehat{\mathcal{M}}) = C_{\widehat{\mathcal{M}}}^*$.

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 detailed¹ explanations.

- An explanation is monotonic iff $C(\pi^*, \hat{\mathcal{M}}) = C_{\hat{\mathcal{M}}}^*$
 $\forall \hat{\mathcal{M}} : \Gamma(\hat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^H) \subset \Gamma(\hat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^H)$.

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.

We will now introduce different kinds of multi-model explanations that can participate in the model reconciliation process, propose algorithms to compute them, and compare and contrast their respective properties. We note that the requirements outlined above are in fact often at odds with each other - an explanation that is very easy to compute may be very hard to comprehend. This (as seen in Table 1) will become clearer in course of this discussion.

A simple way to compute an explanation would be to provide the model differences pertaining to only the actions that are present in the plan that needs to be explained -

A Plan Patch Explanation (PPE) is given by $\mathcal{E}^{MPE} = \Delta_{i \in \{H, R\}} \bigcup_{f \in \{c_a\} \cup \text{pre}(a) \cup \text{eff}^+(a) \cup \text{eff}^-(a) : a \in \pi^* \cap A^i} \tau(f)$.

Clearly, such an explanation is easy to compute and concise by focusing only on plan being explained. However, it may also contain information that need not have been revealed, while at the same time ignoring model differences elsewhere in \mathcal{M}^H that could have contributed to the plan being suboptimal in it. Thus, it is incomplete. An adoption of VAL [Howey *et al.*, 2004] to the MMP setting will be, in fact, a subset of such PPEs, and suffer from the same limitations. An easy way to compute a complete explanation would be to provide the entire model difference to the human -

A Model Patch Explanation (MPE) is given by $\mathcal{E}^{MPE} = \Gamma(\mathcal{M}^R) \Delta \Gamma(\mathcal{M}^H)$.

This is clearly also easy to compute but can be quite large and is hence far from being concise. Thus, in the rest of paper, we will try to minimize the size (and hence increase the comprehensibility) of explanations by searching in the space of models and thereby not exposing information that is not relevant to the plan being explained while still trying to satisfy as many requirements as we can.

A Minimally Complete Explanation (MCE) is the shortest complete explanation, i.e.

$$\mathcal{E}^{MCE} = \arg \min_{\mathcal{E}} |\Gamma(\hat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^H)| \text{ with R1.}$$

The explanation provided before in the Fetch domain is indeed the smallest domain change that may be made to make the given plan optimal in the updated action model, and is thus an example of a minimally complete explanation. An implicit assumptions we make here is that the computation

¹This is a very useful property to have. Doctors, for example, reveal different amount of details of their model to their patients as opposed to their peers. Further, the idea of completeness, i.e. withholding information on other model changes as long as they explain the observed plan, is also quite prevalent in how we deal with similar scenarios ourselves - e.g. progressing from Newtonian physics in high school to Einsteins Laws of Relativity in college.

Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

Table 1: Requirements for different types of explanations.

power (or planning capability) of the human is the same as that of the planner. This means that the human can compute the optimal plan given a planning problem.

Model-space search for MCEs To compute MCEs, we employ A^* search, similar to [Wayllace *et al.*, 2016], in the space of models, as shown in Algorithm 1. Given an MRP, we start off with the initial state $\Gamma(\mathcal{M}^H)$ derived from the human’s expectation of a given planning problem \mathcal{M}^R , and modify it incrementally until we arrive at a planning problem $\hat{\mathcal{M}}$ with $C(\pi^*, \hat{\mathcal{M}}) = C_{\hat{\mathcal{M}}}^*$, i.e. the given plan is explained. Note that the model changes are represented as a set, i.e. there is no sequentiality in the search problem. Also, we assign equal importance to all model corrections. We can easily capture differential importance of model updates by attaching costs to the edit actions λ - the algorithm remains unchanged.

We also employ a selection strategy of successor nodes to speed up search (by overloading the way the priority queue is being popped) by first processing model changes that are relevant to the actions in π_R^* and π_H before the rest.

Proposition 1 The successor selection strategy outlined in Algorithm 1 yields an admissible heuristic for model space search for minimally complete explanations.

Proof Let \mathcal{E} be the MCE for an MRP problem and let \mathcal{E}' be any intermediate explanation found by our search such that $\mathcal{E}' \subset \mathcal{E}$, then the set $\mathcal{E} \setminus \mathcal{E}'$ must contain at least one λ related to actions in the set $\{a \mid a \in \pi_R^* \vee a \in \pi^*\}$ (where π' is the optimal plan for the model $\hat{\mathcal{M}}$ where $\delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\mathcal{M}^H), \mathcal{E}') = \Gamma(\hat{\mathcal{M}})$). To see why this is true, consider an \mathcal{E}' where $|\mathcal{E}'| = |\mathcal{E}| - 1$. If the action in $\mathcal{E} \setminus \mathcal{E}'$ does not belong to either π_R^* or π' then it can not improve the cost of π_R^* in comparison to π' and hence \mathcal{E} can not be the MCE. Similarly we can show that this relation will hold for any size of \mathcal{E}' . We can leverage this knowledge about $\mathcal{E} \setminus \mathcal{E}'$ to create an admissible heuristic that will only consider the relevant changes at any given point of time (by giving very large values to all other changes).

We also note that the optimality criterion is relevant to both the cases where the human expectation is better, or when it is worse, than the plan computed by the planner. This might be counter to intuition, since in the latter case one might expect that just establishing feasibility of a better (than expected optimal) plan would be enough. Unfortunately, this is not the case, as can be easily seen by creating counter-examples where other faulty parts of the human model might disprove the optimality of the plan in the new model -

Proposition 2 If $C(\pi^*, \mathcal{M}^H) < \min_{\pi} C(\pi, \mathcal{M}^H)$, then ensuring feasibility of the plan in the modified planning problem, i.e. $\delta_{\hat{\mathcal{M}}}(\hat{\mathcal{T}}, \pi^*) \models \hat{\mathcal{G}}$, is a necessary but *not* a sufficient condition for $\hat{\mathcal{M}} = \langle \hat{D}, \hat{\mathcal{T}}, \hat{\mathcal{G}} \rangle$ to yield a valid explanation.

Algorithm 1 Search for Minimally Complete Explanations

```

1: procedure MCE-SEARCH
2: Input: MRP  $\langle \pi^*, \langle \mathcal{M}^R, \mathcal{M}^H \rangle \rangle$ 
3: Output: Explanation  $\mathcal{E}^{MCE}$ 
4: Procedure:
5:   fringe  $\leftarrow$  Priority-Queue()
6:   c.list  $\leftarrow \{\}$   $\triangleright$  Closed list
7:    $\pi_R^* \leftarrow \pi^*$   $\triangleright$  Optimal plan being explained
8:    $\pi_H \leftarrow \pi$  such that  $C(\pi, \mathcal{M}^H) = C_{\mathcal{M}^H}^*$   $\triangleright$  Plan expected by human
9:   fringe.push( $\langle \mathcal{M}^H, \{\} \rangle$ , priority = 0)
10:  while True do
11:     $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \leftarrow$  fringe.pop( $\widehat{\mathcal{M}}$ )
12:    if  $C(\pi_R^*, \widehat{\mathcal{M}}) = C_{\widehat{\mathcal{M}}}^*$  then return  $\mathcal{E}$   $\triangleright$  Return  $\mathcal{E}$  if  $\pi_R^*$  optimal in  $\widehat{\mathcal{M}}$ 
13:    else
14:      c.list  $\leftarrow$  c.list  $\cup \widehat{\mathcal{M}}$ 
15:      for  $f \in \Gamma(\widehat{\mathcal{M}}) \setminus \Gamma(\mathcal{M}^R)$  do  $\triangleright$  Models that satisfy condition 1
16:         $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\} \rangle$   $\triangleright$  Removes f from  $\widehat{\mathcal{M}}$ 
17:        if  $\delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$  c.list then
18:          fringe.push( $\langle \delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle$ , c + 1)
19:        for  $f \in \Gamma(\mathcal{M}^R) \setminus \Gamma(\widehat{\mathcal{M}})$  do  $\triangleright$  Models that satisfy condition 2
20:           $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\} \rangle$   $\triangleright$  Adds f to  $\widehat{\mathcal{M}}$ 
21:          if  $\delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$  c.list then
22:            fringe.push( $\langle \delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle$ , c + 1)
23:  procedure PRIORITY_QUEUE.POP( $\widehat{\mathcal{M}}$ )
24:    candidates  $\leftarrow \{ \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c^* \mid c^* = \arg \min_c \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \}$ 
25:    pruned.list  $\leftarrow \{\}$ 
26:     $\pi_H \leftarrow \pi$  such that  $C(\pi, \widehat{\mathcal{M}}) = C_{\widehat{\mathcal{M}}}^*$ 
27:    for  $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \in$  candidates do
28:      if  $\exists a \in \pi_R^* \cup \pi_H$  such that  $\tau^{-1}(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\widehat{\mathcal{M}})) \in \{c_a\} \cup pre(a) \cup$ 
29:         $eff^+(a) \cup eff^-(a)$  then  $\triangleright$  Candidates relevant to  $\pi_R^*$  or  $\pi_H$ 
30:        pruned.list  $\leftarrow$  pruned.list  $\cup \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c$ 
31:    if pruned.list =  $\phi$  then  $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \sim Unif(\text{candidate.list})$ 
32:    else  $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \sim Unif(\text{pruned.list})$ 
    
```

Note that a minimally complete explanation for an MRP can be rendered invalid given further updates to the model. This can be easily demonstrated in our running example in the Fetch domain. Imagine that if, at some point, the human were to find out that the action move also has a precondition (crouched), then the previous robot plan will no longer make sense to the human since now, according to the human's faulty model (being unaware that the tucking action also lowers the robot's torso) the robot would need to do *both* tuck and crouch actions before moving. Consider the following explanation in the Fetch domain instead –

```

Explanation >>
TUCK-has-add-effect-CROUCHED
MOVE_LOC2_LOC1-has-precondition-CROUCHED
    
```

This explanation does not reveal all model differences but at the same time ensures that the robot's plan remains optimal for this problem, irrespective of any other changes to the model, by accounting for all the relevant parts of the model that engendered the plan. It is also the smallest possible among all such explanations. The requirement of monotonicity and minimality brings us to the notion of –

A Minimally Monotonic Explanation (MME) is the shortest explanation that preserves both completeness and monotonicity, given the MRP $\langle \pi^*, \langle \mathcal{M}^R, \mathcal{M}^H \rangle \rangle$, i.e.

$$\mathcal{E}^{MME} = \arg \min_{\mathcal{E}} |\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^H)| \text{ with R1 \& R3.}$$

Algorithm 2 Search for Minimally Monotonic Explanations

```

1: procedure MME-SEARCH
2: Input: MRP  $\langle \pi^*, \langle \mathcal{M}^R, \mathcal{M}^H \rangle \rangle$ 
3: Output: Explanation  $\mathcal{E}^{MME}$ 
4: Procedure:
5:    $\mathcal{E}^{MME} \leftarrow \{\}$ 
6:   fringe  $\leftarrow$  Priority-Queue()
7:   c.list  $\leftarrow \{\}$   $\triangleright$  Closed list
8:   h.list  $\leftarrow \{\}$   $\triangleright$  List of incorrect model changes
9:   fringe.push( $\langle \mathcal{M}^R, \{\} \rangle$ , priority = 0)
10:  while fringe is not empty do
11:     $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \leftarrow$  fringe.pop( $\widehat{\mathcal{M}}$ )
12:    if  $C(\pi^*, \widehat{\mathcal{M}}) > C_{\widehat{\mathcal{M}}}^*$  then
13:      h.list  $\leftarrow$  h.list  $\cup (\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R))$   $\triangleright$  Updating h.list
14:    else
15:      c.list  $\leftarrow$  c.list  $\cup \widehat{\mathcal{M}}$ 
16:      for  $f \in \Gamma(\widehat{\mathcal{M}}) \setminus \Gamma(\mathcal{M}^H)$  do  $\triangleright$  Models that satisfy condition 1
17:         $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\} \rangle$   $\triangleright$  Removes f from  $\widehat{\mathcal{M}}$ 
18:        if  $\delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$  c.list
19:          and  $\nexists S$  s.t.  $(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)) \supseteq S \in$  h.list then  $\triangleright$  Prop 3
20:            fringe.push( $\langle \delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle$ , c + 1)
21:             $\mathcal{E}^{MME} \leftarrow \max_{|\cdot|} \{\mathcal{E}^{MME}, \mathcal{E}\}$ 
22:        for  $f \in \Gamma(\mathcal{M}^H) \setminus \Gamma(\widehat{\mathcal{M}})$  do  $\triangleright$  Models that satisfy condition 2
23:           $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\} \rangle$   $\triangleright$  Adds f from  $\widehat{\mathcal{M}}$ 
24:          if  $\delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$  c.list
25:            and  $\nexists S$  s.t.  $(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)) \supseteq S \in$  h.list then  $\triangleright$  Prop 3
26:              fringe.push( $\langle \delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle$ , c + 1)
27:               $\mathcal{E}^{MME} \leftarrow \max_{|\cdot|} \{\mathcal{E}^{MME}, \mathcal{E}\}$ 
28:     $\mathcal{E}^{MME} \leftarrow (\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)) \setminus \mathcal{E}^{MME}$ 
29:    return  $\mathcal{E}^{MME}$ 
    
```

The last constraint enforces the monotonicity requirement. This means that beyond the model obtained from the minimally monotonic explanation, there do not exist any models which are not explanations of the same MRP, while at the same time making as few changes to the original problem as possible. It follows that this is the largest set of changes that can be done on the planner's planning problem \mathcal{M}^R and still find a model $\widehat{\mathcal{M}}$ where $C(\pi^*, \widehat{\mathcal{M}}) = C_{\widehat{\mathcal{M}}}^*$ - we are going to use this property in the search for MMEs.

Proposition 3 $\mathcal{E}^{MME} = \arg \max_{\mathcal{E}} |\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)|$ such that $\forall \widehat{\mathcal{M}} \Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R) \subseteq \Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)$ it is guarantee to have $C(\pi^*, \widehat{\mathcal{M}}) = C_{\widehat{\mathcal{M}}}^*$.

We also note that an MME solution may not be unique to an MRP problem. This can happen when there are multiple model differences supporting the same causal links in the plan - an MME can get by (i.e. guarantee optimality in the modified model) by only exposing one of them to the human.

Proposition 4 MMEs are not unique, i.e. there might be multiple minimally monotonic solutions to an MRP.

We also note that even though MCEs are an abridged version of an MME, it is easy to see that an MCE may not necessarily be part of an actual MME. This is due to the non-uniqueness of MMEs. Thus, we emphasize -

Proposition 5 An MCE may not be a subset of an MME, but it is always smaller or equal in size, i.e. $|MCE| \leq |MME|$.

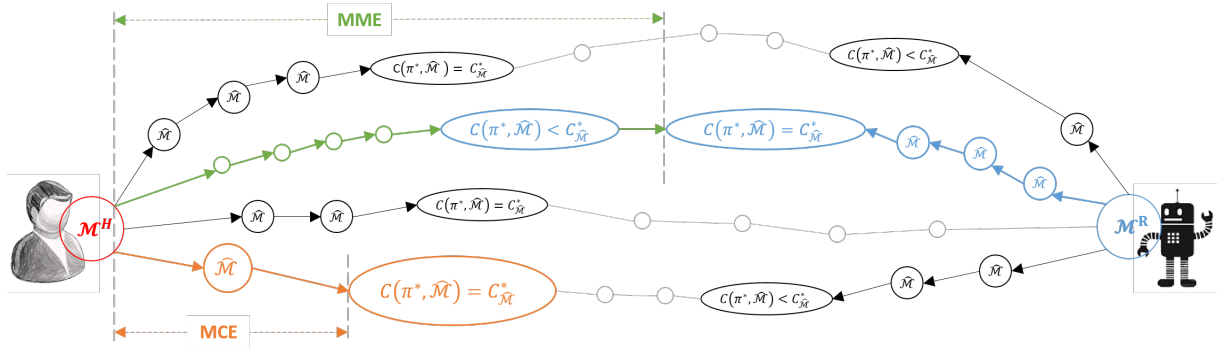


Figure 3: Illustration of model space search for MCE & MME.

Model-space search for MMEs This is similar to the model-space search for MCEs described before, but this time starting from the robot’s model \mathcal{M}^R instead. The goal here is to find the largest set of model changes for which the explicability criterion becomes invalid for the first time (due to either suboptimality or inexecutability). This requires a search over the entire model space, as described in detail in Algorithm 2. We can leverage Proposition 3 to reduce our search space. Starting from \mathcal{M}^R , given a set of model changes \mathcal{E} where $\delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\mathcal{M}^R), \mathcal{E}) = \Gamma(\hat{\mathcal{M}})$ and $C(\pi^*, \hat{\mathcal{M}}) > C_{\hat{\mathcal{M}}}^*$, no superset of \mathcal{E} can lead to an MME solution. In Algorithm 2, we keep track of such unhelpful model changes in the list `h.list`. The variable \mathcal{E}^{MME} keeps track of the current best list of model changes. Whenever we find a new set of model changes where π^* is optimal and is larger than \mathcal{E}^{MME} , we update \mathcal{E}^{MME} with \mathcal{E} . The resulting MME is all the possible model changes that did not appear in \mathcal{E}^{MME} .

Figure 3 contrasts MCE with MME search. MCE search starts from \mathcal{M}^H , computes updates $\hat{\mathcal{M}}$ towards \mathcal{M}^R and returns the first node (indicated in orange) where $C(\pi^*, \hat{\mathcal{M}}) = C_{\hat{\mathcal{M}}}^*$. MME search starts from \mathcal{M}^R and moves towards \mathcal{M}^H . It finds the longest path (indicated in blue) where $C(\pi^*, \hat{\mathcal{M}}) = C_{\hat{\mathcal{M}}}^*$ for all $\hat{\mathcal{M}}$ in the path. The MME (shown in green) is the rest of the path towards \mathcal{M}^H .

Approximate Solution for MCEs

Both MCEs and MMEs may be hard to compute - in the worst case it involves a search over the entire space of model differences. Thus the biggest bottleneck here is the check for optimality of a plan given a new model. A check for necessary or sufficient conditions for optimality, without actually computing optimal plans can be used as a powerful tool to further prune the search tree. In the following section, we thus investigate an approximation to an MCE by employing a few simple proxies to the optimality test. By doing this we lose the completeness guarantee but improve the computability of an explanation. Specifically, we replace the equality test in line 12 of Algorithm 1 by the following rules -

1. $\delta_{\hat{\mathcal{M}}}(\hat{\mathcal{I}}, \pi_R^*) \models \hat{\mathcal{G}}$; and
2. $C(\pi_R^*, \hat{\mathcal{M}}) < C(\pi_R^*, \mathcal{M}^H)$ or $\delta_{\hat{\mathcal{M}}}(\hat{\mathcal{I}}, \pi_H^*) \not\models \hat{\mathcal{G}}$; and
3. Each action contributes at least one causal link to π_R^* .

The first criterion simply ensures that the plan π_R^* originally computed by the planner is actually valid in the new hypothesis model. Criterion (2) requires that this plan has either become better in the new model or at least that the human’s expected plan π_H^* has been disproved. Finally, in Criterion (3), we ensure that for each action $a_i \in \pi_R^*$ there exists an effect p that satisfies the precondition of at least one action a_k (where $a_i \prec a_k$) and there exists no action a_j (where $a_i \prec a_j \prec a_k$) such that $p \in \text{eff}^-(a_j)$.

Proposition 6 Criterion (3) is a necessary condition for optimality of π^* in $\hat{\mathcal{M}}$.

Proof Assume that for an optimal plan π_R^* , there exists an action a_i where criterion (3) is not met. Now we can rewrite π_R^* as $\pi'_R = \langle a_0, a_1, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_n, a_{n+1} \rangle$, where $\text{pre}(a_0) = \phi$ and $\text{eff}^+(a_0) = \{\mathcal{I}\}$ and $\text{pre}(a_{n+1}) = \{\mathcal{G}\}$ and $\text{eff}(a_{n+1}) = \phi$. It is easy to see that $\delta_{\hat{\mathcal{M}}}(\phi, \pi'_R) \models \mathcal{G}$. Now let us consider a cheaper plan $\pi'_R = \langle a_0, a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n, a_{n+1} \rangle$. Since a_i does not contribute any causal links to the original plan π_R^* , we will also have $\delta_{\hat{\mathcal{M}}}(\phi, \pi'_R) \models \mathcal{G}$. This contradicts our original assumption of π_R^* being optimal, hence proved.

4 Empirical Evaluations

Our explanation generation system (as previewed in the Fetch domain) integrates calls to Fast-Downward [Helmert, 2006] for planning, VAL [Howey *et al.*, 2004] for plan validation, and pyperplan [Alkhazraji *et al.*, 2016] for parsing. The results reported here are from experiments run on a 12 core Intel(R) Xeon(R) CPU with an E5-2643 v3@3.40GHz processor and a 64G RAM. The latest version of the code will be available at <https://goo.gl/Bybq7E>. We use three planning domains - BlocksWorld, Logistics and Rover - for our experiments. In order to generate explanations we created the human model by randomly removing parts (preconditions and effects) of the action model. Though the following experiments are only pertaining to action model differences, it does not make any difference at all to the approaches, given the way the state was defined. Also note that these removals, as well as the corresponding model space search, was done in the lifted representation of the domain.

Table 2 In Table 2 we make changes at random to the domains and measure the number of explanations produced and

Domain Name	Problem	MPE (ground truth)		PPE		MME (exact)		MCE (exact w/o heuristic)		MCE (exact with heuristic)		MCE (approximate)	
		size	time	size	time	size	time	size	time	size	time	size	time
BlocksWorld	1	10	n/a	5	n/a	3	1100.8	2	34.7	2	18.9	2	19.8
	2			8		4	585.9	3	178.4	3	126.6	3	118.8
	3			4		5	305.3	2	34.7	2	11.7	2	11.7
	4			7		5	308.6	3	168.3	3	73.3	3	73.0
Rover	1	10	n/a	10	n/a	2	2093.2	2	111.3	2	100.9	2	101.0
	2			10		2	2018.4	2	108.6	2	101.7	2	102.7
	3			10		2	2102.4	2	104.4	2	104.9	2	102.5
	4			9		1	3801.3	1	13.5	1	12.8	1	12.5
Logistics	1	5	n/a	5	n/a	4	13.7	4	73.2	4	73.5	4	63.6
	2			5		4	13.5	4	73.5	4	71.4	4	63.3
	3			5		5	8.6	5	97.9	5	100.4	3	36.4
	4			5		5	8.7	5	99.2	5	95.4	3	36.4

Table 2: Comparison of MCEs and MMEs

$ \mathcal{M}^R \Delta \mathcal{M}^H $	problem-1	problem-2	problem-3	problem-4
3	2.2	18.2	4.7	18.5
5	6.0	109.4	15.4	110.2
7	7.3	600.1	23.3	606.8
10	48.4	6849.9	264.2	6803.6

Table 3: MCE search time for increasing model differences.

BlocksWorld	problem-1	problem-2	problem-3	problem-4
Number of nodes expanded for MME (out of 1024)	128	64	32	32

Table 4: Usefulness of Proposition 3 in pruning MME search.

the time taken (in secs) to produce them, against the ground truth. Observe the gains produced by the heuristic in terms of time spent on each problem. Further, note how close the approximate version of MCEs are to the exact solutions. As expected, MME search is significantly costlier to compute than MCE. However, note that both MCEs and MMEs are *significantly smaller* in size ($\sim 20\%$) than the total model difference (which can be arbitrarily large) in certain domains, further underlining the usefulness of generating minimally complete explanations as opposed to dumping the entire model difference on the human. A general rule of thumb is -

$$|\text{approx. } MCE| \leq |\text{exact. } MCE| < |MME| \ll |MPE|$$

Note that the time required to calculate an MME in the Logistics problems is lower than that for the corresponding MCE. This is because for most of these problems a single change in the planner’s model made the plan be no longer optimal so that the search ended after checking all possible unit changes. In general, closer an MCE is to the total number of changes shorter the MME search would be. Also note how PPE solutions, though much easier to compute, do not have completeness and monotonicity properties, and yet often spans the entire model difference, containing information that are not needed to support the optimality of the given plan.

Table 3 We now increase the number of changes in the human model in BlocksWorld, and illustrate the relative time (in secs) taken to search for exact MCEs in Table 3. As expected there is an exponential increase in the time taken, which can be problematic with even a modest number of model differ-

ences. This further highlights the importance of finding useful approximations to the explanation generation problem.

Table 4 Finally, we demonstrate how Proposition 3 reduces the number of nodes that need to be searched to find MMEs in random problems from the BlocksWorld domain with 10 faults in the human model, as opposed to the total possible 2^{10} models that can be evaluated - equal to the cardinality of the power set of model changes $|\mathcal{P}(\Gamma(\mathcal{M}^R) \Delta \Gamma(\mathcal{M}^H))|$.

5 Conclusions and Future Work

In this paper, we argued that to explain its plans to the human agents in the loop, an AI system needs to explicitly acknowledge that the human may be using a different model than it does. Explanations in this multi-model setting become a process of identifying and reconciling the relevant differences between the models. One immediate future direction is to allow for human’s models that are of different form and/or level of abstraction than the robot’s, so as to allow effective learning of the human’s models (c.f. [Tian *et al.*, 2016; Zhang *et al.*, 2017]), as well as allow for different cognitive abilities of the human using ϵ -optimality or top-K plans [Ribabov *et al.*, 2014] for hypothesis generation.

In cases where the ground truth is not known, the explanation process might also need to consider distributions over relevant models, and iterative refinement of the same via dialog with the human. Work on plan monitoring [Fritz and McIlraith, 2007] can also provide clues to speeding up the search process by providing proxies to the optimality check.

Also note that we insisted that explanations must be compatible with the planner’s model. If this requirement is relaxed, it allows the planner to generate *alternative explanations* that it knows are not true, and thus deceive the human. While endowing the planner with such abilities may warrant significant ethical concerns, we note that the notion of white lies, and especially the relationship between explanations, excuses and lies [Boella *et al.*, 2009] has received very little attention [van Ditmarsch, 2014] and affords a rich set of exciting research problems.

Acknowledgments

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

References

- [Ai-Chang *et al.*, 2004] Mitchell Ai-Chang, John Bresina, Len Charest, Adam Chase, JC-J Hsu, Ari Jonsson, Bob Kanefsky, Paul Morris, Kanna Rajan, Jeffrey Yglesias, et al. Mapgen: mixed-initiative planning and scheduling for the mars exploration rover mission. *IEEE Intelligent Systems*, 19(1):8–12, 2004.
- [Alkhazraji *et al.*, 2016] Yusra Alkhazraji, Matthias Frorath, Markus Grützner, Thomas Liebetaut, Manuela Ortlieb, Jendrik Seipp, Tobias Springenberg, Philip Stahl, Jan Wülfing, Malte Helmert, and Robert Mattmüller. Pyperplan. <https://bitbucket.org/malte/pyperplan>, 2016.
- [Boella *et al.*, 2009] Guido Boella, Jan Broersen, Leendert van der Torre, and Serena Villata. *Representing Excuses in Social Dependence Networks*, pages 365–374. Springer Berlin Heidelberg, 2009.
- [Bryce *et al.*, 2016] Dan Bryce, J. Benton, and Michael W. Boldt. Maintaining evolving domain models. In *IJCAI*, 2016.
- [Ferguson *et al.*, 1996] George Ferguson, James F Allen, Bradford W Miller, et al. Trains-95: Towards a mixed-initiative planning assistant. In *AIPS*, pages 70–77, 1996.
- [Fritz and McIlraith, 2007] Christian Fritz and Sheila A McIlraith. Monitoring plan optimality during execution. In *ICAPS*, pages 144–151, 2007.
- [Göbelbecker *et al.*, 2010] 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.
- [Helmert, 2006] Malte Helmert. The fast downward planning system. *J. Artif. Intell. Res. (JAIR)*, 26:191–246, 2006.
- [Herzig *et al.*, 2014] Andreas Herzig, Viviane Menezes, Leliane Nunes de Barros, and Renata Wassermann. On the revision of planning tasks. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, ECAI’14, 2014.
- [Howey *et al.*, 2004] Richard Howey, Derek Long, and Maria Fox. Val: Automatic plan validation, continuous effects and mixed initiative planning using pddl. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 294–301. IEEE, 2004.
- [Kambhampati, 1990] Subbarao Kambhampati. A classification of plan modification strategies based on coverage and information requirements. In *AAAI 1990 Spring Symposium on Case Based Reasoning*, 1990.
- [Kulkarni *et al.*, 2016] 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.
- [Langley, 2016] Pat Langley. Explainable agency in human-robot interaction. In *AAAI Fall Symposium Series*, 2016.
- [Lombrozo, 2006] Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464 – 470, 2006.
- [Lombrozo, 2012] Tania Lombrozo. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276, 2012.
- [McDermott *et al.*, 1998] Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. Pddl-the planning domain definition language. 1998.
- [Perera *et al.*, 2016] Vittorio Perera, Sai P Selveraj, Stephanie Rosenthal, and Manuela Veloso. Dynamic generation and refinement of robot verbalization. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 212–218. IEEE, 2016.
- [Riabov *et al.*, 2014] Anton Riabov, Shirin Sohrabi, and Octavian Udrea. New algorithms for the top-k planning problem. In *Proceedings of the Scheduling and Planning Applications woRKshop (SPARK) at the 24th International Conference on Automated Planning and Scheduling (ICAPS)*, pages 10–16, 2014.
- [Russell *et al.*, 2003] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.
- [Sohrabi *et al.*, 2011] 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.
- [Tian *et al.*, 2016] Xin Tian, Hankz Hankui Zhuo, and Subbarao Kambhampati. Discovering underlying plans based on distributed representations of actions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1135–1143. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [van Ditmarsch, 2014] Hans van Ditmarsch. *The Ditmarsch Tale of Wonders*, pages 1–12. Springer International Publishing, Cham, 2014.
- [Wayllace *et al.*, 2016] Christabel Wayllace, Ping Hou, William Yeoh, and Tran Cao Son. Goal recognition design with stochastic agent action outcomes. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3279–3285, 2016.
- [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 *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017.