# Excuse My Explanations: Integrating Excuses and Model Reconciliation for Actionable Explanations


Turgay Caglar
*Department of Computer Science*
*Colorado State University*
Fort Collins, USA
turgay.caglar@colostate.edu

Zahra Zahedi
*School of Computing and AI*
*Arizona State University*
Tempe, USA
zzahedi@asu.edu

Sarath Sreedharan
*Department of Computer Science*
*Colorado State University*
Fort Collins, USA
sarath.sreedharan@colostate.edu



***Abstract*—The ability to provide useful and intuitive explanations remains one of the major hurdles to creating robotic systems capable of working effectively with everyday users. In this paper, we consider a popular explanation generation framework for robot task plans, namely model reconciliation, and try to address one of its main drawbacks, namely its inability to generate *actionable explanations*. The current methods for generating model reconciliation focus on generating information that explains why the robot chose a certain behavior over one that was expected by the human. However, the user might also want to understand how they can influence the robot's behavior so it follows the one that was expected from it. Explanations that provide such information are called actionable, and we extend traditional model reconciliation explanations to be actionable by combining them with the existing notion of excuses. We will refer to the resulting explanations as *Actionable Reconciliation Explanations (ARE)*, which explains the robot's decision-making process and suggests how its model might be modified for improved alignment with human expectations. However, as we will see, the generation of ARE requires methods that are distinct from existing model reconciliation and excuse generation methods, and ARE also exhibits properties that are distinct from these earlier methods. We assess our method through computational experiments and user studies and, in the process, also compare it against traditional forms of excuses and model reconciliation explanations.**



***Index Terms*—XAI, Model Reconciliation, Actionable Explanations, Excuses**


## I. INTRODUCTION

The field of Explainable AI, or XAI, has been getting much attention in recent years. While the majority of works in XAI have focused on single-shot decisions like classification [1], [2], there has been increasing interest in developing explanation generation methods for sequential decision-making problems that are better suited for robotics tasks [3], [4]. One of the popular methods in this direction is that of model reconciliation explanations [5], [6]. Given its focus on modeling the user's knowledge and by allowing for the fact that the users might not be aware of all the details of the task or the robotics platform, it is particularly well suited for applications where a lay user may be working with a complex robotics platform. However, these works operate from the assumption that the robot already has the ground truth model of the task, and as such, the robot's plan must be correct and doesn't need to be changed. This, unfortunately, overlooks a core desirable property of explanations that have been identified in the wider XAI literature, namely actionability [7] (also sometimes referred to as algorithmic recourse [8]). An actionable explanation would empower the user, if they wish, to influence or change the system output to the one the user expected or prefers.

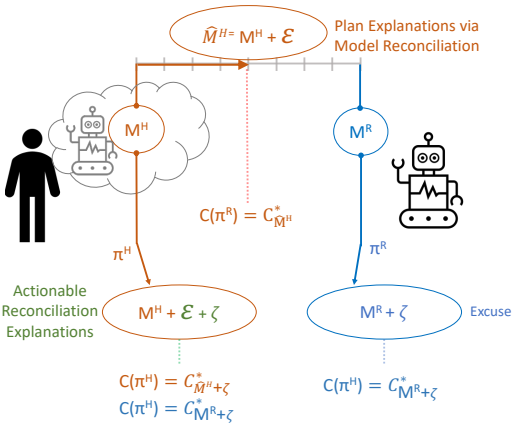


Fig. 1. Schematic representation of model updates for explanation, excuse, and ARE processes. In explanation generation, the human model ($\mathcal{M}^H$) is updated to align more closely with the robot model ($\mathcal{M}^R$), ensuring the robot plan ($\pi^R$) is optimal within the updated human model. During excuse generation, the $\mathcal{M}^R$is updated to ensure that the human's plan ($\pi^H$) is optimal. For ARE following an initial explanation that updates the $\mathcal{M}^H$, the human is also provided with additional model updates that could ensure the optimality of $\pi^H$ in both models.

Excuse generation is an orthogonal approach that looks at how to update a planning model so as to ensure the generation of solutions of some desired property [9], [10]. These methods focus on providing users with information on how to update the task/planning model (most frequently by updating the task setting), such that solving the updated planning model is guaranteed to generate the desired behavior. Unfortunately, excuses on their own do not constitute actionable explanations, as they do not provide any information as to why the system chose the unexpected behavior in the first place [11]. Additionally,


Sarath Sreedharan's research is supported in part by NSF grant NSF 2303019 and other transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program.

the current excuse literature focuses exclusively on scenarios where the user's belief about the robot is equivalent to the true robot model, an assumption that is not met in many scenarios.

In summary, if users only receive model reconciliation explanations generated by the model reconciliation framework, they can understand why the robot's plan is different from what they expect. However, it does not inform them about what needs to be done to achieve the desired behavior. On the other hand, if the robot only provides excuses, users would know what actions to take to get the desired outcome, but they would not understand why the robot initially failed to perform the desired action. In this work, we develop a novel explanation generation method that bridges this gap and generates explanations that help the user understand the reasoning behind the robot's actions and the steps required to achieve their desired behavior.

We show how this new explanatory information, named *Actionable Reconciliation Explanations (ARE)* (Section V), combines the previous approaches' strengths and exhibits a unique set of properties (refer to Fig. 1). We show how generating ARE requires us to develop a novel model-space search algorithm [6] (Section VI). In developing this new method, we also introduce a more general version of the excuse generation problem to the current literature. We evaluate ARE by providing theoretical analyses of these new explanations (Section V), running empirical evaluations to compare how our new algorithm compares to previous ones (Section VII-A), and finally running user studies to establish various effects of the information on the user's end, including, trust, cognitive load, and perceived actionability (Section VII-B). As part of these user studies, we also perform, to the best of our knowledge, the first comparison of excuse and model reconciliation explanations along these dimensions.

## II. Related Work

To start with let's look at the current landscape of relevant related work. A significant portion of previous research on sequential decision-making problems has centered on providing explanations for existing plans. Typically, the inquiries in this area revolve around understanding why a specific plan has been chosen ("Why is this the plan?") or why it was chosen over alternative plans or foils [3], [11]). Some popular explanation methods in this space include model reconciliation [5], [6], causal chain explanations [12], [13], and model restrictions [14]. Excuses, on the other hand, produce information about how a planning model could be updated to produce plans for certain properties. Excuses were originally introduced to identify how an unsolvable model can be made solvable [9]. This can be compared against methods for explaining the unsolvability of the problem [15].

Multiple works have looked at evaluating the effectiveness of model reconciliation explanation. These include basic tests to compare the effectiveness of different forms of model reconciliation explanation and whether people naturally identify model reconciliation explanations (cf. [16]) and what kind of model update information are preferred by users [16]. Early evaluations were focused on simplified robotic tasks [16], but since then they have also been tested in decision-support contexts [17], [18]. Excuses have also been evaluated using user studies [10], but we are unaware of any prior work that has compared the two.

Multiple authors have argued for the need to ensure that explanations generated by AI systems need to be actionable [19]. However, most of the existing works on generating actionable explanations are focused on single-shot decision-making settings [20]. Another related direction of work is that of counterfactual explanations, which were primarily studied in the context of single-shot decision-making. These explanations aim to identify important features behind a decision and how to change them to get different outcomes [21]. Thus, these methods are both identifying the reason for the decision and how to change it [22]. As such, these works are closely in spirit with the kind of methods discussed in this paper.

## III. Background

Our paper will focus on goal-directed deterministic planning problems, i.e., classical planning or STRIPS [23] style planning problems. We can represent such planning problems or equivalently models using tuples of the form $\mathcal{M} = \langle F, A, I, G, C\rangle$, where $\mathcal{M}$ represents a planning model. In this tuple, $F$ is a set of propositional fluents used to define the set of possible unique states within the model. $A$ is the set of actions that can be executed. Each action $a \in A$ is represented as a tuple $\langle pre(a), add(a), del(a)\rangle$, where $pre(a) \subseteq F$, specifies the preconditions for the action. An action $a$ is available in a state $s$ if $pre(a) \subseteq s$. The sets $add(a) \subseteq F$ and $del(a) \subseteq F$ represent the add and delete effects of the action, respectively. We will use the function $\delta_M : 2^F \times A \rightarrow 2^F$ to represent the transition function that captures resulting of executing an action. As such, the result of executing an action $a$ in a given state $s \subseteq F$ is represented as $\delta_M(s, a) = (s \cup add(a)) \setminus del(a)$, where $pre(a) \subseteq s$. $I \subseteq F$ defines the initial state, and $G \subseteq F$ specifies the goal description. Finally, $C$ is a cost function that maps actions to their costs, which are real values.

The solution to this problem is a plan (sequence of actions), $\pi = \langle a_1, a_2, ..., a_n\rangle$, such that $\delta_M(I, \pi) \supseteq G$ (where $\delta_M(I, \pi) = \delta_M((\delta_M(I, a_1), a_2), ...a_n)$). The cost of the plan $\pi$ is $C(\pi) = \sum_{a_i \in \pi} C(a_i)$. We call a plan $\pi$ optimal if no other plan can satisfy the goal description $G$ with less cost. We represent such optimal plans in model $M$ as $\pi^*$, and similarly, the cost for the optimal plan will be $C^*_M$. The original work on excuses focused on unsolvable planning problems, i.e., no action sequence whose execution leads to the goal exists. To capture such cases uniformly, we introduce a new action $a_{-1}$, which has empty preconditions, whose add effects contain the goal description, and has an infinite cost. Thus, a problem is unsolvable if its optimal plan is $\pi_{-1} = \langle a_{-1}\rangle$.

We are interested in setting where the robot's model of the task $\mathcal{M}^R = \langle F, A^R, I^R, G^R, C^R\rangle$ may be different from the human's beliefs about the task $\mathcal{M}^H = \langle F, A^H, I^H, G^H, C^H\rangle$. Keeping with the foundational settings in human-aware planning [24], we will consider a case where the robot is the

primary actor, and the human uses their beliefs to make sense of the robot's actions. For the sake of simplifying the technical discourse, we will assume the two models share the fluent space and action labels (but not necessarily the action definitions). One of the basic problems that could arise here is when the plan followed by the robot ($\pi^R$) is not the same as the one expected by the human ($\pi^H$). To simplify the discussion, we will assume that the expectation mismatch arises from the fact that the human might not think the robot plan is optimal or even valid (per $\mathcal{M}^H$) and the robot might have similar views about the human plan[1]. Both model reconciliation explanations [6] and excuses [9] provide two orthogonal ways of addressing this same problem.

Both these methods use model space search to update either of the two models. In particular, this model space is defined over a space of model parameters $\mathcal{F}^{(F,A)}$, which is defined over the fluents and action labels shared between the two models, and this model parameter set is given as where

$$\mathcal{F}^{(F,A)} = \{\textit{init-has-}f \mid f \in F\} \cup \{\textit{goal-has-}f \mid f \in F\} \cup \bigcup_{a \in A} \{a\textit{-has-pos-prec-}f a\textit{-has-add-}f, a\textit{-has-del-}f \mid f \in F\}.$$

We will now use a parameterization function ($\Gamma(\cdot)$ ) that will map each model to a subset of the model parameters (we will follow the conventions set by [5])

$$\begin{aligned}
\tau_I &= \{\textit{init-has-}f \mid f \in I\} \\
\tau_G &= \{\textit{goal-has-}g \mid g \in G\} \\
\tau_{pre(a)} &= \{a\textit{-has-prec-}f \mid f \in pre(a)\} \\
\tau_{add(a)} &= \{a\textit{-has-add-}f \mid f \in add(a)\} \\
\tau_{del(a)} &= \{a\textit{-has-del-}f \mid f \in del(a)\} \\
\tau_a &= \tau_{pre(a)} \cup \tau_{add(a)} \cup \tau_{del(a)} \\
\tau_A &= \bigcup_{a \in A^{\mathcal{M}}} \tau_a \\
\Gamma(\mathcal{M}) &= \tau_I \cup \tau_G \cup \tau_A
\end{aligned}$$

Similarly, we will use the function $\Gamma^{-1}$ to map a set of model parameters into a model, and we will represent the set of all possible models that can be represented using the model parameters as $\mathbb{M}^{(F,A)}$. Now, under explanation model reconciliation, the goal is to update the human model ($\mathcal{M}^H$), such that the robot plan ($\pi^R$) will be optimal in the updated model. Here, the changes made to the human model align with the robot model; that is, the human is provided with true information about the robot model (that may not have been known previously). We will represent this process of updating the human model as: $\widehat{\mathcal{M}}^H = \mathcal{M}^H + \mathcal{E}$

Where $\widehat{\mathcal{M}}^H$ is the updated human model, '+' is the update operator, and $\mathcal{E}$ is the model information (i.e., the explanation). Note that the use of the '+' operator is one of convenience since the update may provide the human with information about the robot model they previously didn't know (as such, add new parameters to $\Gamma(\mathcal{M}^H)$), and possibly correct misconception they may have about the robot model (as such, remove parameters from $\Gamma(\mathcal{M}^H)$). We will use the cost function $\mathcal{C}^{\mathcal{E}}$ to return the non-negative cost of providing an explanation (i.e., a piece of model information) and will refer to an explanation set with the lowest cost as a minimally complete explanation or *MCE*. Henceforth, unless specified, we use the term explanation to refer to MCE. Now, an excuse, on the other hand, is to update the robot model such that the solutions generated meet certain criteria. Since we will be introducing a more general notion of excuses, we will provide a detailed definition of excuses in Section V.

[1]It is worth noting that most of the following discussion still holds if the human is not performing optimal planning

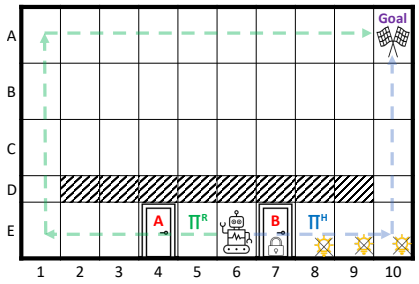


Fig. 2. Motivational Example: The robot, initially positioned at E6 within a small room bounded by walls and two doors (unknown to the human Door B is locked), is tasked by a human to reach the goal at A10. A continuous wall extends from D2 to D9, forming an impassable barrier for the robot. The corridor between E8 and E10 is dark due to the lights being turned off. The human's anticipated route for the robot is depicted with blue arrows ($\pi^H$), while the robot's actual path is indicated by green arrows ($\pi^R$).

## IV. Running Example

To illustrate the distinctions between excuses, explanations, and our method, we introduce a scenario featuring a robot navigating a small grid world, as depicted in Fig. 2. Initially positioned in a small room at E6, the robot is tasked by a human to reach a specified goal at A10 via the shortest possible route. The environment includes a wall marked by diagonal lines extending from D2 to D9, which the robot cannot cross. There are two exit doors: Door A at E4 and Door B at E7. The corridor between E8 and E10 is dark due to the lights being switched off, which the human is aware of. The optimal plan expected by the human, denoted as $\pi^H$ in blue, involves the robot exiting via Door B to directly reach the goal. The robot, on the other hand, exits through Door A and follows a seemingly longer path to the goal ($\pi^R$ denoted in green). Here, the human is unaware of the fact that the robot can't open locked doors, that door B is locked, and that the robot can't navigate through dark corridors.

In the presence of this discrepancy, if the user were to ask, "Why didn't you just go through door B to reach the goal?" the robot could respond either with an excuse or an explanation.
**Explanation** - *I can't open the doors that are locked and Door B is locked*
**Excuse** - *I could have followed $\pi^H$ if Door B was unlocked and the lights at cells E8, E9, and E10 is switched on.*
Now, an explanation provides information as to why the robot

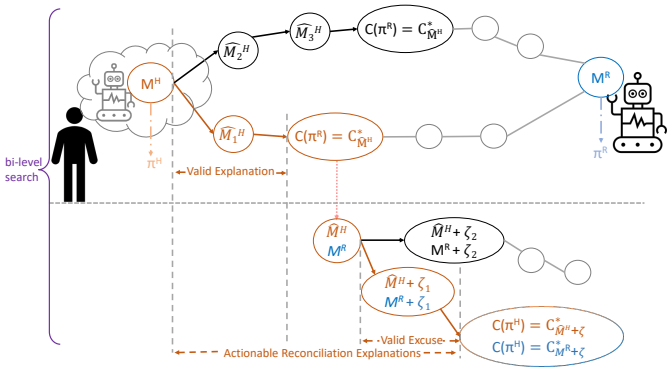


Fig. 3. Illustration of the Bi-Level Search Algorithm for identifying ARE. This process is initiated by locating a valid explanation, followed by a subsequent search phase that seeks a valid excuse, taking into account both the robot's model and the human model updated with the initial valid explanation.

chose its behavior but not how to correct it, while excuses, on the other hand, provide information on how to allow generation of $\pi^H$, but no justification for its choice of $\pi^R$ .

The minimal ARE for the setting would be as follows

**ARE** - *I can't open the doors that are locked, I can't move in the dark, and door B is locked. I could have followed $\pi^H$ if Door B was unlocked and the lights at cells E8, E9, and E10 is switched on.*

Note that ARE tries to perform functions of both explanations and excuses; however, it contains information not contained in either. We will provide a more detailed discussion of the reason why ARE takes this form and how to generate them in the following sections.

## V. Actionable Reconciliation Explanations

Before discussing and analyzing our newly proposed explanatory information, let's revisit the problem of excuse generation and start with a more general definition.

*Definition 1: For a target model $\mathcal{M} = \langle F, A, I, G\rangle$ and a target plan set $\Pi$, and **general excuse generation problem** is defined by the tuple $\mathcal{P}^\zeta = \langle \mathcal{M}, \mathbb{E}^\zeta, \mathcal{C}^\zeta, \Pi\rangle$, where $\mathbb{E}^\zeta \subseteq \mathcal{F}^{(F,A)}$ is the set of allowed model edits, and $\mathcal{C}^\zeta$ is the cost associated with each model edit. A solution to an excuse generation problem is an* excuse $\zeta \subseteq \mathbb{E}^\zeta$, *such that for the updated model $\mathcal{M}^\zeta = \mathcal{M} + \zeta$, there exists at least a plan $\pi \in \Pi$, that is optimal in $\mathcal{M}^\zeta$.*

The original excuse generation work [9] focused on converting an unsolvable planning problem into a solvable one by changing the initial state. We can derive this specific instantiation of the problem from the general problem by setting the target plan set, to the set of all action sequences not containing $a_{-1}$. In our case, our target model is the robot model $\mathcal{M}^R$, the target plan set is a singleton set containing $\pi^H$, and the model update set $\mathbb{E}^\zeta$ contains all the changes to the task and the robot that can be viably made, and $\mathcal{C}^\zeta$ reflect the cost of making that change. In the context of our running example, the model edits include both changing the environment (unlocking Door B) and upgrading the robot (adding sensors), and the identified minimal excuse includes edits of both kinds.

Now, moving over to the problem at hand, the underlying explanatory problem is one shared by multiple earlier works (cf. [5], [17], [25]). In its most general form, the problem is represented as a tuple $\mathcal{P}^\mathcal{E} = \langle \mathcal{M}^H, \mathcal{M}^R, \mathcal{C}^\mathcal{E}, \pi^R, \pi^H\rangle$. Even though these earlier works look at the same problem, they propose generating different kinds of explanations for different objectives. For example, works on model reconciliation [24] have introduced both MCE for one-shot interactions and MME for longitudinal ones. Similarly, works on explicable planning [26] consider changing the robot plans to better align with human expectations. We add to this growing body of literature by considering a novel explanation form that becomes more appropriate when we relax a core assumption shared by all these previous works, namely that the robot model is not changeable. In particular, the explanatory information generated under this new method consists of two sets of information: a) a set of information about the robot model that is aimed at explaining why the robot chose $\pi^R$, b) a set of model updates that if performed could allow the robot to follow $\pi^H$ instead. Thus, the explanation here includes the reason for the system's choice ($\pi^R$) and a set of actionable recourse the user could potentially follow to generate the behavior they expected ($\pi^H$). We will again leverage the notion of model edits to define such explanations. Specifically, we formalize the notion of Actionable Reconciliation Explanations ($ARE$) as follows:

*Definition 2: For an actionable explanation problem $\mathcal{P} = \langle \mathcal{M}^H, \mathcal{M}^R, \mathbb{E}^\zeta, \mathcal{C}^\mathcal{E}, \mathcal{C}^\zeta, \pi^R, \pi^H\rangle$, the $ARE$ is given as a tuple $\Xi = \langle \mathcal{E}, \zeta\rangle$, such that it meets the following conditions:*

C1 $\zeta \subseteq \mathbb{E}$
C2 $\pi^R$ *is optimal in the model* $\mathcal{M}^H + \mathcal{E}$
C3 $\pi^H$ *is optimal in both the models* $\mathcal{M}^H + \mathcal{E} + \zeta$ *and* $\mathcal{M}^R + \zeta$

Under this definition, $\mathcal{E}$ corresponds to the explanation component and $\zeta$ to the excuse (C1 merely states that the excuse can only include valid model updates). Condition C2 requires that $\mathcal{E}$ helps ensure the optimality of the robot plan in the updated human model, and condition C3 requires that the application of excuse ensures the optimality of the human plan in both the robot model and the updated human model you obtain after incorporating the explanation.

Our natural next step would be to attribute a notion of minimality to ARE. A natural option would be to minimize $\mathcal{C}^\zeta(\zeta) + \mathcal{C}^\mathcal{E}(\mathcal{E})$. However, revisiting the motivational example shows how this could be an issue. Note how the minimal ARE provided includes more information than a simple union of the explanation and excuse information. If we were to simply return the union of the two, the user would understand why the robot would want to unlock Door B, but not why the lights need to be switched on. For the second piece of information, it needs to know that the robot can't go through dark rooms. In this example, a naive approach to simply minimizing the total information provided could lead to the human thinking that the excuse includes more information than is required. So, instead, we will use a multi-tiered notion of minimality. In particular, for a given ARE, we are looking for the cheapest explanation, for which we can generate an excuse that is perceived as

minimal by the human. Note that this definition centers on human perception. As they do not know the robot model, it is very hard for humans to judge what a minimal explanation is. However, as in the example, once an explanation is given, the human can try to estimate if the excuse includes superfluous information. To formally capture this notion, we will start by defining a minimal excuse for an explanation.

*Definition 3: For a given problem $\mathcal{P}$, a set $\zeta \subseteq \mathbb{E}$ is said to be minimal for $\mathcal{E}$, if $\langle \mathcal{E}, \zeta \rangle$ is a valid ARE for $\mathcal{P}$ and there exists no other excuse $\zeta' \subseteq \mathbb{E}$, such that $\pi^H$ is optimal for $\mathcal{M}^H + \zeta'$ and $\mathcal{C}^\zeta(\zeta') < \mathcal{C}^\zeta(\zeta)$.*

A minimal ARE is one with a minimal explanation for which a minimal excuse is possible, or more formally

*Definition 4: For a given problem $\mathcal{P}$, an ARE $\Xi = \langle \mathcal{E}, \zeta \rangle$ is considered minimal if $\zeta$ is minimal for $\mathcal{E}$ and there exists no $\Xi' = \langle \mathcal{E}', \zeta' \rangle$, such that $\mathcal{C}^\mathcal{E}(\mathcal{E}') < \mathcal{C}^\mathcal{E}(\mathcal{E})$ and $\zeta'$ is minimal for $\mathcal{E}'$.*

Now, with the definitions in place, we can see some basic properties of ARE

*Proposition 1: For a given problem $\mathcal{P}$, there might not exist a valid ARE.*

This result is in stark contrast with the model reconciliation explanation, where one could always show one exists. Here, this property primarily arises from the requirements for a valid excuse, and one can show this fact holds trivially when $\mathbb{E}$ is empty (i.e., the robot model cannot be updated in line with the original assumption of the model reconciliation work).

Returning to a property that was earlier hinted at in the example, the total cost of a minimal ARE might be higher than the total cost of a minimal explanation and excuse.

*Proposition 2: For a given problem $\mathcal{P}$, let $\Xi = \langle \mathcal{E}, \zeta \rangle$ be a minimal ARE, now let $\mathcal{E}^*$ be the MCE for $\mathcal{P}^\mathcal{E}$, and $\zeta^*$ be the minimal excuse for $\mathcal{P}^\zeta$, then*

$$\mathcal{C}^\zeta(\zeta) + \mathcal{C}^\mathcal{E}(\mathcal{E}) \geq \mathcal{C}^\zeta(\zeta^*) + \mathcal{C}^\mathcal{E}(\mathcal{E}^*)$$

The proof of this proposition can be established rather directly. Firstly, we can see from Definitions 2 and 4 that the excuse and explanation components of any ARE are valid explanations and excuses. Since $\mathcal{C}^\mathcal{E}(\mathcal{E}^*)$ and $\mathcal{C}^\zeta(\zeta^*)$ are lower bounds on explanations and excuses, their sum must also be a lower bound of the total ARE cost. This establishes the $\geq$ relationship. We can see that it's neither always equal nor greater in the general case through construction. The running example already shows a case where the cost ARE is higher, and in the evaluation, we will see cases where they are equal. This eliminates any chances of establishing a more precise relationship without further constraining the problem.

## VI. Generating ARE

To generate ARE, we will be using a bi-level search (refer to Fig. 3), termed ARE-search. The outer-search will be similar to the explanation generation process described for generating MCE (cf. [6]). In the traditional MCE, the goal check merely checks whether the identified model updates constitute a valid model reconciliation explanation (with guarantees of minimality provided by the choice of the search). Now, in our case, we will run an additional goal check when this condition is met. In particular, we check if a valid minimal excuse exists for the corresponding explanation $\mathcal{E}$.

**Algorithm 1** Algorithm of the excuse generation pseudo-code procedure that is called by the outer model-space search when it identifies a valid explanation.

**procedure** MINIMAL EXCUSE SEARCH
  *Input*: $\mathcal{M}^R, \hat{\mathcal{M}}^H, \mathbb{E}, \pi^H, \mathcal{C}^\zeta$, where $\hat{\mathcal{M}}^H = \mathcal{M}^H + \mathcal{E}$
  *Output*: Excuse set $\zeta$ and *min_excuse_found* flag
  fringe $\leftarrow$ `Priority_Queue()`
  *min_excuse_found* $\leftarrow$ `False`
  $\hat{\zeta} \leftarrow \{\}$
  fringe.push($\hat{\zeta}$, priority $= 0$)
  *min_excuse_cost* $\leftarrow \infty$
  **while** fringe is not empty **do**
    $\hat{\zeta} \leftarrow$ fringe.pop()
    **if** $\mathcal{C}^\zeta(\hat{\zeta}) \leq$ *min_excuse_cost* **then**
      **if** $\pi^H$ is optimal in $\hat{\mathcal{M}}^H + \hat{\zeta}$ **then**
        *min_excuse_cost* $\leftarrow \mathcal{C}^\zeta(\hat{\zeta})$
        **if** $\pi^H$ is optimal in $\mathcal{M}^R + \hat{\zeta}$ **then**
          *min_excuse_found* $\leftarrow$ `True`
          **return** $\hat{\zeta}$, *min_excuse_found*
    **for** $\forall e \in \mathbb{E} \setminus \hat{\zeta}$ **do**
      $\hat{\zeta} \leftarrow \hat{\zeta} \cup \{e\}$
      fringe.push($\hat{\zeta}, \mathcal{C}^\zeta(\hat{\zeta}) + h^\zeta(\hat{\zeta})$)
    **return** $\hat{\zeta}$, *min_excuse_found*

The algorithm for finding such an excuse is sketched in Algorithm VI. It takes as input the updated human model (obtained by applying an explanation of the human model), the original robot model, the allowed edits, the human plan, and the cost function. The algorithm internally makes use of a modified A* search [27] with an admissible heuristic ($h^\zeta$). It searches over the space of excuses (corresponding to subsets of $\mathbb{E}$). It will only test an excuse if its cost is lower than or equal to the previously found *min_excuse_cost*. This *min_excuse_cost* is initially set to infinity (or more practically to a large value). The *min_excuse_cost* variable gets set to a specific value as soon as you find an excuse that makes the plan optimal in the human model. If the same excuse makes it optimal in the robot model, then it's returned as the required minimal excuse, else it continues searching. The successors for each expanded excuses involves new excuse set formed by adding previously missing model edits from $\mathbb{E}$, and the node value is set as the sum of it's cost and heuristic value (as per the conventions of A* search).

## VII. Evaluation

### A. Experiments on IPC Domains

First, we determine our proposed algorithm's computational characteristics, particularly the time taken on standard planning benchmarks, and compare them against the original model reconciliation explanation (referred to henceforth as simply explanation) and excuse generation methods[2]. The time taken is an interesting metric of comparison because we are

[2]All code and data for the experiments can be found in https://github.com/cglrtrgy/ActionableExplanations

TABLE I

COMPARISON OF PERFORMANCE METRICS ACROSS THREE APPROACHES: EXCUSE, EXPLANATION, AND ARE. FOR EACH DOMAIN, METRICS ARE AVERAGED ACROSS 5 PROBLEM INSTANCES, WITH STANDARD DEVIATIONS PROVIDED. $|\pi^R|$ AND $|\pi^H|$ REPRESENT THE AVERAGE LENGTHS OF ROBOT AND HUMAN PLANS, RESPECTIVELY. $|\zeta|$, $|\mathcal{E}|$, AND $|\zeta + \mathcal{E}|$ DENOTE THE LENGTHS OF THE EXCUSE, EXPLANATION, AND ARE. THE 'TIME(S)' COLUMN INDICATES THE COMPUTATION TIME (IN SECONDS).

| Domains | $\|\pi^R\|$ | $\|\pi^H\|$ | Excuse | | Explanation | | ARE | |
|---|---|---|---|---|---|---|---|---|
| | | | $\|\zeta\|$ | Time(s) | $\|\mathcal{E}\|$ | Time(s) | $\|\zeta + \mathcal{E}\|$ | Time(s) |
| Logistics | 31.0 ± 11.0 | 15 ± 4.5 | 2.8 ± 1.3 | 5.2 ± 5.7 | 2.0 ± 0.0 | 22.1 ± 46.4 | 4.8 ± 1.3 | 16.2 ± 16.7 |
| Depots | 19.0 ± 7.6 | 8.2 ± 3.6 | 3.6 ± 1.3 | 4.5 ± 5.3 | 3.8 ± 0.4 | 137 ± 249.1 | 7.4 ± 1.5 | 153 ± 264.7 |
| Freecell | 12.0 ± 5.6 | 5.0 ± 1.6 | 2.6 ± 1.9 | 16.4 ± 34.4 | 2.8 ± 0.4 | 18.6 ± 19.3 | 5.4 ± 2.2 | 45.6 ± 72.4 |
| Rovers | 13.0 ± 5.5 | 7.8 ± 3.1 | 2.6 ± 0.9 | 6.6 ± 11.5 | 2.0 ± 0.0 | 1.4 ± 2.0 | 4.6 ± 0.9 | 8.1 ± 12.7 |
| Satellite | 16.2 ± 2.4 | 14.2 ± 2.2 | 2.6 ± 1.1 | 9.2 ± 13.7 | 1.8 ± 0.8 | 11.9 ± 17.3 | 4.4 ± 1.8 | 27.7 ± 32.6 |

TABLE II

THIS TABLE FOCUSES ON PROBLEM INSTANCES WHERE THE LENGTH OF ARE IS GUARANTEED TO BE GREATER THAN THE TOTAL LENGTH OF EXCUSE AND EXPLANATION WHEN COMPUTED INDIVIDUALLY. SEE TABLE I FOR A DETAILED DESCRIPTION OF THE METRICS.

| Domains | $\|\pi^R\|$ | $\|\pi^H\|$ | Excuse | | Explanation | | ARE | |
|---|---|---|---|---|---|---|---|---|
| | | | $\|\zeta\|$ | Time(s) | $\|\mathcal{E}\|$ | Time(s) | $\|\zeta + \mathcal{E}\|$ | Time(s) |
| Blocks | 12.0 ± 2.4 | 8.8 ± 1.1 | 2.8 ± 0.8 | 2.9 ± 3.8 | 1.0 ± 0.0 | 0.3 ± 0.0 | 4.6 ± 1.1 | 5.2 ± 3.0 |
| Zenotravel | 13.6 ± 3.2 | 12.0 ± 3.1 | 2.0 ± 0.0 | 1.7 ± 1.8 | 1.0 ± 0.0 | 13.5 ± 15.2 | 4.0 ± 0.0 | 24.8 ± 26.7 |
| Logistics | 31.0 ± 11.0 | 13.0 ± 5.4 | 2.4 ± 0.9 | 3.4 ± 6.5 | 2.0 ± 0.0 | 21.9 ± 42.4 | 5.2 ± 0.4 | 38.8 ± 37.5 |

introducing a new class of more complex search algorithms (given the multi-level structure) than the ones previously considered in the literature.

*a) Setting :* For the experiments, we will assign uniform unit costs for all model updates (both for explanations and excuse model edits) and use a blind heuristic for all model-space searches. We considered two settings, one where we tested on the original benchmark domains and then a set where we know that the cost of ARE will be higher than the total cost for explanation and excuses. For the first set, the robot model consisted of the original IPC domains and instances, and the human model was generated by randomly deleting some static predicates from the original domain. The allowed model updates for excuses includes adding a subset of predicates to the initial state. We considered five domains, and for each domain, we considered five instances of increasing size (25 total problems).

We also evaluated our approach on problems where ARE is guaranteed to be longer than individual excuses or explanations. As such, we expect the search effort to be significantly higher in these problems. To create such problem instances, we updated each domain by adding a duplicate set of actions that allows for shorter plans by removing some of the preconditions in the human model. As such, to explain away the use of these actions, the system only needs to point to one precondition in one of these actions. However, to ensure the optimality of human plans that use these actions, a minimal ARE will involve explanations that include all relevant preconditions. The human models were generated as before, and the model updates for excuses were again similar to the last setting. For this setting, we considered three domains and five instances each (15 total problems). All evaluations were performed on a computer with 16GB RAM and an Apple M1 3.2GHz CPU.

*b) Results:* Tables I and II present the time taken and solution size for the three methods, i.e., excuses, explanations, and ARE. Since all costs are unit costs, we report costs simply as the size. Table I focuses on the first setting where ARE could simply be a union of the explanation and excuse. The primary thing to note here is the fact that even though the algorithm for ARE is more complex than a simple model search involved in excuse and explanation generation, the time taken by ARE is comparable to the sum of time taken for excuse and explanation. This is even true in Table II, where we are explicitly considering problems where the solution size for ARE is larger than the sum of excuse and explanation in isolation. Please note that the reason for the large standard deviations is that we are summarizing results across planning instances of different sizes. Refer to Appendix E to see the results for each domain-problem instance pair.

### B. Human Subject Experiments

Next, using a between-subjects study, we compared our method against just excuses and model reconciliation explanations. Each participant was randomly assigned to one of three conditions: Excuse only (EXC), Explanation only (EXP), or Our Method: ARE. The experiment followed IRB approved protocols, and an overview of the study setup and the robot behavior can be seen in the attached video[3].

*a) Study design and task:* Within each condition, participants were presented with two scenarios involving a robot assigned to perform a task. The sequence of these scenarios was randomized per participant. The participants were briefed on the robot task, and then they watched a recorded video of the robot performing the task (for example, Fig. 4), where they saw the robot performing a seemingly suboptimal plan. Following each task, a question is raised about its behavior, and the robot responds according to the condition (EXC, EXP, or ARE). Details of the users' questions and the robot's responses are provided in Fig. 5. One of the tasks produces an $ARE$ that is the union of explanation and excuses (Task 1 in Fig. 5)

[3]Video link: https://youtu.be/JpPO0RVHtkg

TABLE III

POST-HOC TEST RESULTS FOR THE DEPENDENT VARIABLES CLARITY AND ACTIONABILITY, THE ONLY VARIABLES FOR WHICH ANOVA SHOWED SIGNIFICANT DIFFERENCES. * DENOTES SIGNIFICANT COMPARISONS WHERE $p_{tukey} < 0.05$, BASED ON TUKEY'S POST-HOC CORRECTION.

| DV | Comparison | Mean Difference | SE | t | $p_{tukey}$ |
|---|---|---|---|---|---|
| Clarity | ARE - Excuse-only | 0.218 | 0.062 | 3.534 | 0.002* |
| | ARE - Explanation-only | 0.140 | 0.061 | 2.310 | 0.057 |
| | Excuse-only - Explanation-only | -0.078 | 0.061 | -1.272 | 0.413 |
| Actionability | ARE - Excuse-only | 0.179 | 0.059 | 3.025 | 0.008* |
| | ARE - Explanation-only | 0.317 | 0.058 | 5.455 | <.001* |
| | Excuse-only - Explanation-only | 0.138 | 0.059 | 2.358 | 0.051 |

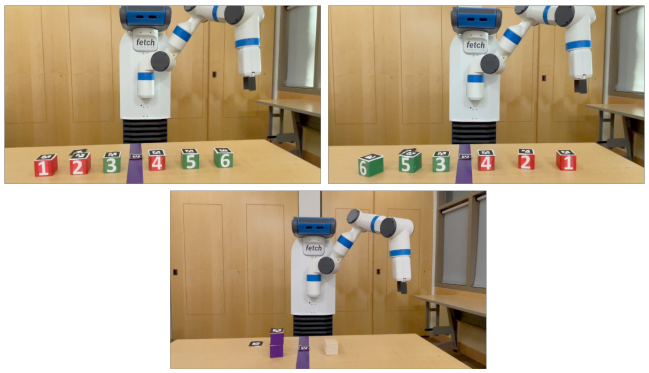

Fig. 4. Images captured from an actual robot in two different task scenarios. Top images illustrate the first task, where the robot groups blocks by color on opposite table sides using minimal steps. The top left image displays the initial setup, while the top right shows the post-task arrangement, highlighting that the robot's plan of swapping two red blocks (1 and 2) on the left with two green blocks (5 and 6) on the right is seemingly sub-optimal (instead of swapping the one green block three on the left with red block four on the right). The bottom image presents the initial setup of an unsolvable task, where the robot was to stack a clear block on top of the purple blocks.

and in the other it contains more information (Task 2 in Fig. 5). Participants then evaluated the robot's response in terms of *satisfaction*, *clarity*, and their perceived ability to make the robot behave the way they wanted (*actionability*). After completing both tasks, participants responded to questions regarding their *trust* in the robot and their perceived *workload*. The survey concluded with demographic questions.

*b) Participants and recruitment:* We recruited 93 participants through the Prolific [28]. We excluded data from three participants who failed an attention check. The analysis proceeded with the following number of participants: 29 in the 'EXC' condition, 31 in the 'EXP' condition, and 30 in the ARE condition.

On average, participants completed the survey in eight minutes and were compensated $3.50 for their participation. In our study, 54% of participants identified as men, 41% as women, and 4% as non-binary. The largest age group, comprising 41% of participants, was between 25-34 years old. Regarding education, 5% of participants held a college degree or higher, 33% had a high school diploma or its equivalent, and 16% had obtained a graduate degree. Within the participant pool, 14% had a degree in Computer Science.

*c) Dependent variables:* We evaluated participants in three conditions across two tasks. We were particularly interested in analyzing the following dependent variables (DV):

**Task 1 User's Question:** *Why did you choose not to swap the green block on the left side (block 3) with the red block on the right side (block 4)?*
**Task 1 Robot's Responses:**

- **EXC:** If the red block 4 was lighter, I could have swapped the green block on the left side (block 3) with the red block on the right side (block 4).
- **EXP:** Red block 4 is heavy. I cannot pick up heavy objects.
- **ARE:** I cannot pick up the heavy blocks. Red block 4 is heavy. If red block 4 was lighter, I could have performed only one swap.

**Task 2 User's Question:** *Why can't you solve this task?*
**Task 2 Robot's Responses:**

- **EXC:** If the clear block was on the purple divider, I could have performed the task.
- **EXP:** I cannot perform the task because I cannot pick up the clear block from the table.
- **ARE:** I cannot perform the task because I cannot pick up the clear block from the table. I can pick up clear blocks that are placed on the purple divider. If the clear block was on the purple divider, I could have performed the task.

Fig. 5. User questions and robot responses


Fig. 6. Box Plots of Conditions by Dependent Variables. 'X' represents the mean and the line represents median.

- *Satisfaction*: This measures how satisfied participants were with the robot's response to the tasks they were given.
- *Clarity*: This assesses how clear the participants found the robot's responses.
- *Actionability*: This assesses participants' understanding of how to make the robot perform the tasks in the way they wanted.
- *Trust*: The trust the user placed on the robot, as measured through Muir questionnaire [29].
- *Workload*: The workload, especially cognitive load, imposed by each condition, as measured by NASA TLX questionnaire [30].

*d) Hypotheses:* For the human subject experiments, the following hypotheses were tested:

- **H1** ARE condition will result in higher satisfaction compared to the EXC and EXP conditions.
- **H2** ARE condition will result in higher clarity compared to the EXC and EXP conditions.
- **H3** ARE condition will result in higher perceived actionability compared to the EXC and EXP conditions.
- **H4** ARE condition will result in higher trust compared to the EXC and EXP conditions.
- **H5** No significant difference in workload is expected between ARE and EXC or EXP conditions.

*e) Results:* A one-way ANOVA was performed to compare the effect of the robot's response according to the condition (EXC, EXP, or ARE) on each DV. We present the results in relation to each hypothesis, incorporating the mean and standard deviation (SD) values for further insight.

**H1:** While ANOVA did not show a significant difference in *satisfaction* scores between conditions ($F(2, 177) = 2.053, p = 0.131, \omega^2 = 0.012$), the mean satisfaction score was highest in the ARE condition ($M = 0.414, SD = 0.362$) compared to the EXC ($M = 0.302, SD = 0.330$) and EXP ($M = 0.312, SD = 0.311$) conditions. This trend aligns with H1 but does not reach statistical significance. **H2:** ANOVA revealed a significant difference in clarity scores between conditions ($F(2, 177) = 6.441, p = 0.002, \omega^2 = 0.057$). Post hoc Tukey's test indicated that the ARE condition ($M = 0.678, SD = 0.339$) resulted in *higher* clarity scores with statistically significant $p$ value compared to the EXC condition ($M = 0.460, SD = 0.342$, $p = 0.002$). However, the difference between the ARE and EXP conditions did not reach statistical significance ($p = 0.057$). Additionally, the ARE condition had the highest clarity score among all conditions, followed by EXP ($M = 0.538, SD = 0.324$). These results partially support H2. **H3:** The ARE condition led to statistically significant higher *actionability* scores ($M = 0.653, SD = 0.325$) compared to both the EXC condition ($M = 0.474, SD = 0.336, p = 0.008$) and the EXP condition ($M = 0.336, SD = 0.301, p < 0.001$), supporting H3. **H4:** For *trust*, ANOVA did not show a significant difference between conditions ($F(2, 87) = 0.544, p = 0.583, \omega^2 = 0.000$). While the ARE condition ($M = 0.271, SD = 0.212$) had the highest *trust* scores compared to EXC ($M = 0.270, SD = 0.236$) and EXP ($M = 0.220, SD = 0.201$), the differences were not statistically significant. **H5:** ANOVA results showed no significant difference in *workload* scores between conditions ($F(2, 87) = 1.501, p = 0.229, \omega^2 = 0.011$), supporting H5. Table III presents the results of all pairwise comparisons performed as part of Tukey's HSD test for DVs that showed significant differences in the ANOVA analysis. Descriptive statistics for all DVs across different conditions are displayed in box plots in Fig. 6 and Table VI in Appendix C.

*f) Discussion:* ANOVA and post-hoc test results revealed that the *actionability* score for ARE was higher with statistically significant p values than for both EXC and EXP conditions. Additionally, we saw that the *clarity* score for the ARE condition was higher, with a statistically significant p-value, than that for the EXC condition. These results align with our primary hypothesis regarding the utility of ARE. Specifically, it provides users with clear information that not only helps them understand the rationale behind the robot's behavior but also informs them about the necessary changes required to elicit the expected behavior from the robot. These results suggest that ARE is, in fact, improving the "actionability" of the model reconciliation explanation while not reducing its explanatory power. It is also interesting to note that the users in fact found ARE to be more actionable than 'EXC'. This might point to the fact that while the 'EXC' condition may list the required changes, a lack of rationale about why these changes are needed may leave the user confused. As such, reducing the user's ability to utilize this information correctly.

Interestingly, ANOVA did not reveal significant differences in *satisfaction*, *trust*, or *workload* scores across the conditions. One possible explanation for the lack of variance in *satisfaction* is that users' perceptions of robots would influence their expectations and, in turn, how satisfied they are with the robot. Since our study focused on a block placement task, participants may have perceived the robot as performing adequately, regardless of the explanation provided. In terms of *workload*, it is worth noting that the users observed the same robot performing the same tasks across all three conditions. The only difference across conditions was the explanatory information provided. In each case, the user is expected to look at the same behavior, try to make sense of it, and then evaluate the explanation. This cross-condition similarity of what the user needs to do, combined with the relatively simple nature of the explanations involved, might explain the similarity in *workload*. Finally, in the case of *trust*, it is well established that user trust is directly related to their willingness to take risks and to be vulnerable when using the AI system [31]. As such, the similarity of the robot behavior and a lack of any significant risk on the user's end from using the robot could also explain the similarity in perceived *trust*.

## VIII. Conclusion

This paper introduces a novel form of model reconciliation that not only explains why the system chose a behavior but also how the user could potentially change it. In doing so, we not only model reconciliation explanations actionable but also combine them with the existing methods of 'excuse' generation. This new form of explanation is validated through user studies, which shows its advantages over explanations and excuses. In the study, we also perform a comparison between excuses and explanations, marking a first in this field of research. Additionally, we apply our approach to IPC domains to study the computational characteristics of our proposed algorithm for generating ARE. Future work will involve conducting semi-structured interviews with users to delve deeper into the factors they consider to assess excuses, explanations, and ARE's, thereby further enriching our understanding of these methods.

## REFERENCES


[1] X. Wang and M. Yin, "Effects of explanations in ai-assisted decision making: Principles and comparisons," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–36, 2022.

[2] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 2019, pp. 563–574.

[3] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable ai planning and decision making," *arXiv preprint arXiv:2002.11697*, 2020.

[4] ——, "Human-aware planning revisited: A tale of three models," in *IJCAI-ECAI XAI/ICAPS XAIP Workshops*, vol. 10, 2018.

[5] S. Sreedharan, T. Chakraborti, and S. Kambhampati, "Foundations of explanations as model reconciliation," *Artificial Intelligence*, vol. 301, p. 103558, 2021.

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

[7] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[8] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[9] M. Göbelbecker, T. Keller, P. Eyerich, M. Brenner, and B. Nebel, "Coming Up With Good Excuses: What to do When no Plan Can be Found," in *ICAPS*, 2010.

[10] M. Diehl, T. Chakraborti, and K. Ramirez-Amaro, "Guided demonstrations using automated excuse generation," *arXiv preprint arXiv:2311.18355*, 2023.

[11] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial intelligence*, 2019.

[12] P. Bercher, S. Biundo, T. Geier, T. Hoernle, F. Nothdurft, F. Richter, and B. Schattenberg, "Plan, repair, execute, explain—how planning helps to assemble your home theater," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 24, 2014, pp. 386–394.

[13] S. Sreedharan, C. Muise, and S. Kambhampati, "Generalizing action justification and causal links to policies," in *ICAPS*. AAAI Press, 2023, pp. 417–426.

[14] B. Krarup, S. Krivic, D. Magazzeni, D. Long, M. Cashmore, and D. E. Smith, "Contrastive explanations of plans through model restrictions," *Journal of Artificial Intelligence Research*, vol. 72, pp. 533–612, 2021.

[15] S. Sreedharan, S. Srivastava, D. Smith, and S. Kambhampati, "Why can't you do that hal? explaining unsolvability of planning tasks," in *International Joint Conference on Artificial Intelligence*, 2019.

[16] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan explanations as model reconciliation–an empirical study," in *HRI*. Ieee, 2019, pp. 258–266.

[17] K. Valmeekam, S. Sreedharan, S. Sengupta, and S. Kambhampati, "Radar-x: An interactive mixed initiative planning interface pairing contrastive explanations and revised plan suggestions," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 32, 2022, pp. 508–517.

[18] S. Grover, S. Sengupta, T. Chakraborti, A. P. Mishra, and S. Kambhampati, "Radar: automated task planning for proactive decision support," *Human–Computer Interaction*, vol. 35, no. 5-6, pp. 387–412, 2020.

[19] B. Kim and F. Doshi-Velez, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[20] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[21] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai)," in *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28*. Springer, 2020, pp. 163–178.

[22] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[23] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3-4, pp. 189–208, 1971.

[24] S. Sreedharan, A. Kulkarni, and S. Kambhampati, *Explainable Human-AI Interaction: A Planning Perspective*. Springer Nature, 2022.

[25] S. Sreedharan, T. Chakraborti, C. Muise, and S. Kambhampati, "Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2518–2526.

[26] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, "Plan explicability and predictability for robot task planning," in *ICRA*. IEEE, 2017, pp. 1313–1320.

[27] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[28] S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[29] B. M. Muir, "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[30] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[31] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of management review*, vol. 23, no. 3, pp. 393–404, 1998.

## APPENDIX A
## USER STUDY QUESTIONS

In Table IV, we provide the specific questions and statements used to assess participant responses across various metrics in our user study. This table categorizes the survey items under the dependent variables they are intended to measure, including *Trust*, *Workload*, *Satisfaction*, *Clarity*, and *Actionability*. Each item is accompanied by a corresponding 7-point Likert scale.

TABLE IV
OVERVIEW OF QUESTIONS AND STATEMENTS ASSOCIATED WITH EACH DEPENDENT VARIABLE.

| Dependent Variables | User Study Questions/Statements | 7-point Scale |
|---|---|---|
| Satisfaction | I am satisfied with the answer to my question provided by the robot. | 1-Low, 7-High |
| Clarity | I found the robot's answer regarding its behavior to be clear. | 1-Low, 7-High |
| Actionability | I understand how to make this robot perform this task in the way I wanted. | 1-Low, 7-High |
| Trust | To what extent can the robot's behavior be predicted from moment to moment? | 1-Low, 7-High |
| | To what extent can you count on the robot to do its job? | 1-Low, 7-High |
| | What degree of faith do you have that the robot will be able to cope with similar situations in the future? | 1-Low, 7-High |
| | Overall, how much do you trust the robot? | 1-Low, 7-High |
| Workload | How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving? | 1-Low, 7-High |
| | How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious? | 1-Low, 7-High |
| | How much time pressure did you feel due to the pace at which tasks occurred? Was the pace slow and leisurely or rapid and frantic? | 1-Low, 7-High |
| | How successful were you in accomplishing the goals of the task? How satisfied were you with your performance? | 1-Good, 7-Poor |
| | How hard did you have to work to accomplish your level of performance? | 1-Low, 7-High |
| | How did you feel during the task? | 1-Low, 7-High |

## APPENDIX B
## DEMOGRAPHICS BY CONDITIONS

Table V presents detailed breakdown of the demographic distribution of participants according to the response condition they were assigned to: Excuse-only, Explanation-only, and ARE. Each condition column, denoted by 'n=', represents the total number of participants in that specific group, providing a basis for the subsequent percentage calculations within each demographic category. These categories encompass Gender, Age, Education of participants, and whether participants have Computer Science (CS) Background or not, with their respective percentages summing up to 100% for each condition. The 'Overall' column, marked as 'N=90', aggregates the data across all conditions, offering a comprehensive view of the entire study's demographic landscape.

TABLE V
PARTICIPANT DEMOGRAPHIC DISTRIBUTION BY CONDITIONS

| Demographics | Categories | Excuse (n=29) | Explanation (n=31) | ARE (n=30) | Overall (N=90) |
|---|---|---|---|---|---|
| Gender | Woman | 27.59% | 48.39% | 46.67% | 41.11% |
| | Man | 65.52% | 48.39% | 50% | 54.44% |
| | Non-binary | 6.90% | 3.23% | 3.33% | 4.44% |
| Age | 18 -24 | 13.79% | 16.13% | 16.67% | 15.56% |
| | 25 -34 | 41.38% | 41.94% | 40% | 41.11% |
| | 35 - 44 | 17.24% | 25.81% | 10% | 17.78% |
| | 45 - 54 | 13.79% | 9.68% | 23.33% | 15.56% |
| | 55 - 64 | 10.34% | 6.45% | 10% | 8.89% |
| | 65+ | 3.45% | 0% | 0% | 1.11% |
| Education | High school | 34.48% | 32.26% | 33.33% | 33.33% |
| | College | 51.72% | 48.39% | 53.33% | 51.11% |
| | Graduate | 13.79% | 19.35% | 13.33% | 15.56% |
| CS Background | No | 82.76% | 90.32% | 83.33% | 85.56% |
| | Yes | 17.24% | 9.68% | 16.67% | 14.44% |

## APPENDIX C
## DESCRIPTIVE STATISTICS FOR DEPENDENT VARIABLES BY CONDITIONS

In Table VI, we present the mean and standard deviation (SD) values for various dependent variables under three different conditions: Excuse-only, Explanation-only, and ARE.

TABLE VI
MEAN AND STD FOR DEPENDENT VARIABLES BY CONDITIONS

| Dependent Variable | Excuse | Explanation | ARE |
|---|---|---|---|
| Satisfaction | $0.302 \pm 0.330$ | $0.312 \pm 0.311$ | $0.414 \pm 0.362$ |
| Clarity | $0.460 \pm 0.342$ | $0.538 \pm 0.324$ | $0.678 \pm 0.339$ |
| Actionability | $0.474 \pm 0.336$ | $0.336 \pm 0.301$ | $0.653 \pm 0.325$ |
| Trust | $0.270 \pm 0.236$ | $0.220 \pm 0.201$ | $0.271 \pm 0.212$ |
| Workload | $0.321 \pm 0.109$ | $0.272 \pm 0.107$ | $0.288 \pm 0.118$ |

## APPENDIX D
## TWO-TAILED T-TEST AND TWO ONE-SIDED T-TESTS (TOST) RESULTS

In Table VII, we present detailed statistical analyses to assess the equivalence of trust and workload across conditions.

TABLE VII
STATISTICAL ANALYSES INCLUDING THE TWO-TAILED T-TEST RESULTS (DISPLAYED IN THE 'STATISTIC' COLUMN UNDER 'T-TEST') AND THE TWO ONE-SIDED T-TESTS (TOST) RESULTS (PRESENTED IN THE 'STATISTIC' COLUMN, UNDER 'UPPER BOUND' AND 'LOWER BOUND' SECTIONS). P-VALUES MARKED WITH "*" DENOTE SIGNIFICANCE AT $p < 0.05$

| Dependent Variable | Statistic | Excuse vs Explanation | | | Excuse vs Action | | | Excuse vs Explanation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | t | df | p | t | df | p | t | df | p |
| Trust | T-Test | 0.880 | 58 | 0.382 | 0.012 | 57 | 0.990 | 0.953 | 59 | 0.345 |
| | Upper bound | 1.074 | 58 | 0.144 | 0.204 | 57 | 0.419 | 1.148 | 59 | 0.128 |
| | Lower bound | 0.686 | 58 | 0.752 | -0.180 | 57 | 0.429 | 0.758 | 59 | 0.774 |
| Workload | T-Test | 1.761 | 58 | 0.083 | -1.109 | 57 | 0.272 | 0.563 | 59 | 0.576 |
| | Upper bound | 1.955 | 58 | 0.028* | -0.917 | 57 | 0.819 | 0.758 | 59 | 0.226 |
| | Lower bound | 1.568 | 58 | 0.939 | -1.301 | 57 | 0.099 | 0.367 | 59 | 0.643 |

## APPENDIX E
## PRE-AGGREGATION RESULTS FOR DOMAIN-PROBLEM INSTANCE PAIRS

Table VIII and Table IX provide detailed breakdowns of the aggregated results presented in Table I and Table II, respectively.

TABLE VIII

COMPARISON OF PERFORMANCE METRICS ACROSS THREE APPROACHES: EXCUSE, EXPLANATION, AND ARE. FOR EACH DOMAIN AND PROBLEM INSTANCE PAIRS. $|\pi^R|$ AND $|\pi^H|$ REPRESENT THE AVERAGE LENGTHS OF ROBOT AND HUMAN PLANS, RESPECTIVELY. $|\zeta|$, $|\mathcal{E}|$, AND $|\zeta + \mathcal{E}|$ DENOTE THE LENGTHS OF THE EXCUSE, EXPLANATION, AND ARE. THE 'TIME(S)' COLUMN INDICATES THE COMPUTATION TIME (IN SECONDS). SUMMARIZED DATA CAN BE FOUND IN TABLE I

| Domains | Problem | $\|\pi^R\|$ | $\|\pi^H\|$ | Excuse | | Explanation | | ARE | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $\|\zeta\|$ | Time(s) | $\|\mathcal{E}\|$ | Time(s) | $\|\zeta + \mathcal{E}\|$ | Time(s) |
| **Logistics** | p1 | 20 | 10 | 1 | 0.14 | 2 | 0.46 | 3 | 0.99 |
| **Logistics** | p2 | 19 | 11 | 2 | 0.19 | 2 | 0.48 | 4 | 1.02 |
| **Logistics** | p3 | 36 | 17 | 4 | 11.83 | 2 | 1.62 | 6 | 27.66 |
| **Logistics** | p4 | 36 | 16 | 3 | 3.11 | 2 | 2.74 | 5 | 12.41 |
| **Logistics** | p5 | 44 | 21 | 4 | 10.6 | 2 | 105.1 | 6 | 38.85 |
| **Depots** | p1 | 10 | 4 | 2 | 2.71 | 4 | 3.60 | 6 | 4.71 |
| **Depots** | p2 | 15 | 6 | 3 | 0.69 | 4 | 4.32 | 7 | 5.34 |
| **Depots** | p3 | 27 | 12 | 5 | 2.47 | 4 | 96.94 | 9 | 110.42 |
| **Depots** | p4 | 16 | 7 | 3 | 13.78 | 3 | 3.52 | 6 | 24.67 |
| **Depots** | p5 | 27 | 12 | 5 | 2.68 | 4 | 576.68 | 9 | 620.11 |
| **Freecell** | p1 | 8 | 5 | 1 | 0.24 | 2 | 3.33 | 3 | 3.04 |
| **Freecell** | p2 | 14 | 6 | 2 | 1.72 | 3 | 26.04 | 5 | 34.13 |
| **Freecell** | p3 | 8 | 4 | 2 | 1.02 | 3 | 7.08 | 5 | 8.73 |
| **Freecell** | p4 | 21 | 7 | 6 | 78.01 | 3 | 49.2 | 9 | 173.37 |
| **Freecell** | p5 | 9 | 3 | 2 | 1.09 | 3 | 7.12 | 5 | 8.87 |
| **Rovers** | p1 | 10 | 5 | 2 | 0.55 | 2 | 0.49 | 4 | 1.47 |
| **Rovers** | p2 | 14 | 8 | 3 | 3.52 | 2 | 0.52 | 5 | 4.53 |
| **Rovers** | p3 | 11 | 7 | 2 | 0.51 | 2 | 0.49 | 4 | 1.18 |
| **Rovers** | p4 | 8 | 6 | 2 | 1.34 | 2 | 0.47 | 4 | 2.55 |
| **Rovers** | p5 | 22 | 13 | 4 | 27.1 | 2 | 5.06 | 6 | 30.73 |
| **Satellite** | p1 | 17 | 16 | 1 | 0.16 | 1 | 0.54 | 2 | 1.27 |
| **Satellite** | p2 | 15 | 13 | 2 | 1.06 | 2 | 2.48 | 4 | 5.98 |
| **Satellite** | p3 | 20 | 17 | 4 | 7.40 | 3 | 41.31 | 7 | 62.1 |
| **Satellite** | p4 | 15 | 12 | 3 | 33.14 | 2 | 13.71 | 5 | 64.53 |
| **Satellite** | p5 | 14 | 13 | 3 | 4.23 | 1 | 1.49 | 4 | 4.67 |

TABLE IX

COMPARISON OF PERFORMANCE METRICS ACROSS THREE APPROACHES: EXCUSE, EXPLANATION, AND ARE. $|\pi^R|$ AND $|\pi^H|$ REPRESENT THE AVERAGE LENGTHS OF ROBOT AND HUMAN PLANS, RESPECTIVELY. $|\zeta|$, $|\mathcal{E}|$, AND $|\zeta + \mathcal{E}|$ DENOTE THE LENGTHS OF THE EXCUSE, EXPLANATION, AND ARE. PLEASE NOTE THAT THE LENGTH OF ARE IS GREATER THAN THE TOTAL LENGTH OF EXCUSE AND EXPLANATION WHEN COMPUTED INDIVIDUALLY. THE 'TIME(S)' COLUMN INDICATES THE COMPUTATION TIME (IN SECONDS). SUMMARIZED DATA CAN BE FOUND IN TABLE II

| Domains | Problem | $\|\pi^R\|$ | $\|\pi^H\|$ | Excuse | | Explanation | | ARE | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $\|\zeta\|$ | Time(s) | $\|\mathcal{E}\|$ | Time(s) | $\|\zeta + \mathcal{E}\|$ | Time(s) |
| **Blocks** | p1 | 10 | 8 | 2 | 0.89 | 1 | 0.35 | 4 | 4.00 |
| **Blocks** | p2 | 16 | 10 | 4 | 10.22 | 1 | 0.36 | 6 | 9.55 |
| **Blocks** | p3 | 12 | 10 | 3 | 1.70 | 1 | 0.35 | 5 | 5.00 |
| **Blocks** | p4 | 10 | 8 | 2 | 0.27 | 1 | 0.33 | 3 | 1.38 |
| **Blocks** | p5 | 12 | 8 | 3 | 1.56 | 1 | 0.34 | 5 | 6.27 |
| **Zenotravel** | p1 | 11 | 9 | 2 | 0.20 | 1 | 0.29 | 4 | 2.02 |
| **Zenotravel** | p2 | 14 | 12 | 2 | 1.20 | 1 | 7.18 | 4 | 14.39 |
| **Zenotravel** | p3 | 10 | 9 | 2 | 0.19 | 1 | 0.29 | 4 | 1.94 |
| **Zenotravel** | p4 | 15 | 14 | 2 | 2.69 | 1 | 32.08 | 4 | 44.76 |
| **Zenotravel** | p5 | 18 | 16 | 2 | 4.34 | 1 | 27.43 | 4 | 60.78 |
| **Logistics** | p1 | 20 | 8 | 2 | 0.40 | 2 | 0.93 | 5 | 8.79 |
| **Logistics** | p2 | 19 | 8 | 2 | 0.47 | 2 | 0.88 | 5 | 8.59 |
| **Logistics** | p3 | 36 | 14 | 2 | 0.72 | 2 | 3.06 | 5 | 39.72 |
| **Logistics** | p4 | 36 | 14 | 2 | 0.48 | 2 | 7.05 | 5 | 100.39 |
| **Logistics** | p5 | 44 | 21 | 4 | 15.12 | 2 | 97.64 | 6 | 36.47 |

## APPENDIX F
## LIMITATION

Like all research, our study is subject to certain limitations. One of them lies in our capacity to interpret the findings. Despite our concerted efforts to engage in thorough discussions among co-authors and to draw insights from related studies in the field, the complexity of data interpretation remains a challenge.

Additionally, while we endeavored to include a broad spectrum of participants, achieving a perfectly representative sample is inherently difficult. There remains the possibility that not all participant groups were adequately represented in our study. Consequently, we advocate for future research to prioritize inclusivity and diversity among participant groups to enhance the generalizability of findings.

Our investigation was also confined to a limited selection of IPC domains, which may not encompass the full range of potential applications. Expanding the scope to include a wider variety of domains could provide a more comprehensive evaluation of the phenomena under study.

In our human subject experiments, we did not ask users to make the changes required to achieve the expected plans. Instead, we used a metric to assess the perceived actionability of our method. We did not require a separate test to measure actual actionability since our algorithm guarantees the soundness of the excuses generated. We also did not include additional experiments for explainability metrics, such as simulability, as the utility of model reconciliation as an explanation framework has been extensively studied (cf. [16]). However, future studies could explore the metrics like simulability and the ability of the users to carry out the changes recommended by the method.

Lastly, the nature of our study does not allow for the assessment of long-term implications associated with the types of information we examined. To gain a deeper understanding of the enduring effects, conducting longitudinal studies in real-world settings would be invaluable. Such research could offer significant insights into how these information types influence interactions and perceptions over extended periods.