On the Computational Complexity of Model Reconciliations

Sarath Sreedharan[1] , **Pascal Bercher**[2] and **Subbarao Kambhampati**[1]
[1]School of Computing & AI, ASU
[2]The Australian National University
sarath.sreedharan@colostate.edu, pascal.bercher@anu.edu.au, rao@asu.edu

Abstract

Model-reconciliation explanation is a popular framework for generating explanations for planning problems. While the framework has been extended to multiple settings since its introduction for classical planning problems, there is little agreement on the computational complexity of generating minimal model reconciliation explanations in the basic setting. In this paper, we address this lacuna by introducing a decision-version of the model-reconciliation explanation generation problem and we show that it is Σ_2^P-complete.

1 Introduction

The problem of generating intuitive and concise explanations for plans generated by AI agents has been receiving a lot of attention in recent years [Hoffmann and Magazzeni, 2019; Fox *et al.*, 2017; Chakraborti *et al.*, 2020]. Model reconciliation [Sreedharan *et al.*, 2021] is an explanation generation framework popular in planning that frames explanation as a process of bringing human's expectation about the robot closer to the robot's behavior by updating their beliefs about the robot model, i.e., the human's theory of mind about the robot. While the original work does present a clear definition of the central explanatory challenge, namely, identifying a set of model updates of minimal length, we are unaware of any works that successfully establish the computational complexity of this problem. This is particularly unfortunate since solution strategies for model-reconciliation explanation tend to rely on extremely general, but computationally expensive model-space search. An exact characterization of the complexity of the problem could help us determine whether such methods are truly warranted in the case of model-reconciliation explanation.

In this paper, we focus on the basic problem studied by Chakraborti *et al.* [2017]. We formalize a decision-version of the minimal explanation problem studied there and show that this decision problem is in fact Σ_2^p-complete. The proof will focus on mapping the explanation problem to that of establishing the satisfiability of a particular subclass of quantified boolean formulas. Specifically, one where the quantification is restricted to a set of existentially quantified variables followed by a set of universally quantified variables. The reduction for the membership proof will leverage the universal quantification in the formula to capture the optimality test that is a central part of the explanation generation problem and the existentially quantified variables to capture the explanation. While in the case of the hardness proof, we will use the optimality test to capture the universal quantification and the explanation to capture the existential one.

2 Background

We start by providing basic definitions for planning problems and model-reconciliation and will end the section by providing a brief introduction to some of the relevant complexity classes we will be considering in this paper.

2.1 Classical Planning

We will focus on deterministic, goal-directed planning problems represented in the STRIPS planning problem formalism [Geffner and Bonet, 2013]. Specifically a STRIPS planning problem (henceforth referred to as planning model) may be formally represented by a tuple $\mathcal{M} = \langle F^{\mathcal{M}}, A^{\mathcal{M}}, \delta^{\mathcal{M}}, I^{\mathcal{M}}, G^{\mathcal{M}} \rangle$, where $F^{\mathcal{M}}$ is a set of proposition fluents that define the state space associated with the planning problem (i.e., $S^{\mathcal{M}} = 2^{F^{\mathcal{M}}}$); $A^{\mathcal{M}}$ is a set of action names; $\delta^{\mathcal{M}} = \langle pre_+^{\mathcal{M}}, pre_-^{\mathcal{M}}, add^{\mathcal{M}}, del^{\mathcal{M}} \rangle$ provides the functions that map a given action name to its positive precondition, negative precondition, add effects and delete effects, such that

$$pre_+^{\mathcal{M}} : A \to 2^{F^{\mathcal{M}}} \qquad add^{\mathcal{M}} : A \to 2^{F^{\mathcal{M}}}$$
$$pre_-^{\mathcal{M}} : A \to 2^{F^{\mathcal{M}}} \qquad del^{\mathcal{M}} : A \to 2^{F^{\mathcal{M}}}$$

Note that when we refer to an "action" we usually mean its name $a \in A^{\mathcal{M}}$, not its precondition and effect tuple $(pre_+^{\mathcal{M}}(a), pre_-^{\mathcal{M}}(a), add^{\mathcal{M}}(a), del^{\mathcal{M}}(a))$. $I^{\mathcal{M}} \in S^{\mathcal{M}}$ is the initial state from which the agent is trying to achieve the goal; and $G^{\mathcal{M}} \subseteq F^{\mathcal{M}}$ is the goal specification, where $S_G^{\mathcal{M}} = \{s \in S^{\mathcal{M}} \mid s \supseteq G^{\mathcal{M}}\}$ is the set of goal states.

An action $a \in A^{\mathcal{M}}$ is called executable in a state $s \in S^{\mathcal{M}}$ when its preconditions are satisfied by the current state, formally represented by $exe(a, s, \mathcal{M}) = pre_+^{\mathcal{M}}(a) \subseteq s$ and $s \cap pre_-^{\mathcal{M}}(a) = \emptyset$. The effect of executing an action is represented by the function $\gamma(a, s, \mathcal{M})$ defined as:

$$= \begin{cases} (s \setminus del^{\mathcal{M}}(a)) \cup add^{\mathcal{M}}(a), & \text{if } exe(a, s, \mathcal{M}) = true \\ undefined & \text{otherwise} \end{cases}$$

Executability of sequences of actions is defined by subsequent action application.

The functions *exe* and γ take a model $\mathcal{M}$ as an additional parameter. The reason for this is that we will make changes to specific actions and thus require to consider action executability and its state transition function in different models.

A (possibly empty) sequence of actions $\pi = \langle a_1, ..., a_k \rangle$ is called a *solution* (also: *plan*) if it is executable in the initial state and results into a goal state, i.e., $\gamma(\pi, I^{\mathcal{M}}, \mathcal{M}) = \gamma(a_k, \gamma(...(\gamma(a_1, I^{\mathcal{M}}, \mathcal{M}))...) \supseteq G^{\mathcal{M}}$. Additionally, each action and by extension the plan can be associated with a cost. In this paper, we will specifically focus on models where each action has a unit cost. Thus the cost of the plan, denoted by $C(\pi)$, is equal to the length of the plan. As usual for heuristics, we use a star to denote optimality. A plan π^* is said to be a optimal for a model $\mathcal{M}$, if

$$\nexists\, \pi' \text{ with } \gamma(\pi', I^{\mathcal{M}}, \mathcal{M}) \supseteq G^{\mathcal{M}}, \text{ such that } C(\pi') < C(\pi^*).$$

We will use the notation $C^*_{\mathcal{M}}$ to capture the cost of an optimal plan for $\mathcal{M}$, i.e., the length of any shortest solution for $\mathcal{M}$.

2.2 Encoding Planning Problems as SAT

A popular way of solving planning problems, particularly when there exists a planning horizon (say T), is to encode it as propositional satisfiability problems (SAT) [Kautz *et al.*, 1996]. The most common encoding for the problem uses propositional variables to capture whether a fluent is true at each possible time step and whether an action was executed at a time step. If $\mathcal{M}$ is the planning model, then we would have $T \times (|F^{\mathcal{M}}| + |A^{\mathcal{M}}|)$ variables. The encoding has three important classes for clauses (a) clauses that describe a component of a model, i.e, initial state, goal, or action definition, (b) explanatory frame axioms that enforce the requirement that any change in variable value should correspond to the execution of an action that could have caused the change and finally (c) clauses to enforce the fact that concurrent action execution is not possible. For example, let $a_i \in A^{\mathcal{M}}$ and $p \in add^{\mathcal{M}}(a)$, then as part of class (a) of clauses you would have clause of the form $a_i^t \Rightarrow p^{t+1}$, for all time steps t. Similarly, an example of an explanatory clause for p would be

$$\neg p^t \wedge p^{t+1} \Rightarrow \bigvee \{a_i^t | p \in add^{\mathcal{M}}(a)\}$$

Effectively the clause asserts that p could only have been turned true if one of these actions was executed. Finally, we assert that the actions cannot be executed concurrently using the clause

$$\bigwedge_{a \in A} (a^t \Rightarrow \bigwedge_{a_j \in A, a_j \neq a} \neg a_j^t)$$

2.3 Model Reconciliation Explanation

The basic setting of model reconciliation explanation consists of an autonomous agent (henceforth referred to as the Robot R), using its model $\mathcal{M}^R$ to come up with its plans. There is a human observer, who assumes the robot uses a model $\mathcal{M}_h^R$ (i.e., the robot model assumed by the human) and is trying to make sense of the robot's plan. If the model $\mathcal{M}_h^R$ is different from $\mathcal{M}^R$, the human may be confused by the robot's choice to follow some (robot's) plan π_R, as it might appear suboptimal or even invalid (i.e., not executable or not achieving all goals) according to the model $\mathcal{M}_h^R$. Now the goal of the model reconciliation is to resolve this confusion by providing the human with information on how the human's assumed model differs from the actual (robot) model.

Before we can formally define a model reconciliation explanation problem, we need to define a model parameterization function. We will follow a formalization slightly different from those used by Chakraborti *et al.* [2017] and Sreedharan *et al.* [2021] and define the formalization around a space of models that share the same fluent space and action names.

Definition 1. *Given a set of propositions F and a set of action names A, let $\mathbb{M}^{(F,A)}$ be the **space of models** that can be defined over F and A, i.e., $\forall \mathcal{M} \in \mathbb{M}^{(F,A)}$ there exist δ, I, and G, such that $\mathcal{M} = \langle F, A, \delta, I, G \rangle$ is a planning model.*

Now each model from a given model space can be uniquely identified by a so-called model parameterization function, based on the definition set up by Chakraborti *et al.* [2017].

Definition 2. *The **model parameterization** function $\Gamma : \mathbb{M}^{(F,A)} \rightarrow 2^{\mathcal{F}^{(F,A)}}$ for a given space of models $\mathbb{M}^{(F,A)}$, maps a model from $\mathbb{M}^{(F,A)}$ to a subset of propositions $\mathcal{F}^{(F,A)}$ (henceforth referred to as model parameters), where*

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

For a model $\mathcal{M} = \langle F^{\mathcal{M}}, A^{\mathcal{M}}, \delta^{\mathcal{M}}, I^{\mathcal{M}}, G^{\mathcal{M}} \rangle$, the parameterization function $\Gamma(\mathcal{M})$ is defined by

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

Note that $\Gamma : \mathbb{M}^{(F,A)} \rightarrow 2^{\mathcal{F}^{(F,A)}}$ is a bijective mapping and we will use the function Γ^{-1} to identify the model in $\mathbb{M}^{(F,A)}$, corresponding to a specific subset of $\mathcal{F}^{(F,A)}$.

Definition 3. *A **model reconciliation explanation problem** is defined by the tuple $\mathcal{P}^{MRE} = \langle \mathcal{M}^R, \mathcal{M}_h^R, \pi_R^* \rangle$, where $\mathcal{M}^R = \langle F^{\mathcal{M}^R}, A^{\mathcal{M}^R}, \delta^{\mathcal{M}^R}, I^{\mathcal{M}^R}, G^{\mathcal{M}^R} \rangle$ is a model that the robot is using in its decision-making; $\mathcal{M}_h^R = \langle F^{\mathcal{M}_h^R}, A^{\mathcal{M}_h^R}, I^{\mathcal{M}_h^R}, G^{\mathcal{M}_h^R} \rangle$ is the model the human observer is associating with the robot; and π_R^* is the robot's plan to be explained. We demand that the models share the same fluents and action names $F^{\mathcal{M}^R} = F^{\mathcal{M}_h^R}$, $A^{\mathcal{M}^R} = A^{\mathcal{M}_h^R}$, and that π_R^* is optimal in the model $\mathcal{M}^R$.*

Note that the requirement of using identical action names and fluents is not a restriction. Using the same action name set is a canonical requirement as we assume that the only confusion that might exist is due to misaligned action definitions, i.e., the human observer might not have a perfect understanding of an action's preconditions and effects, but does know which action is being observed. Requiring identical fluent sets is just for convenience, but not a restriction either, since we can always define this shared/identical fluent set as the union of the individual ones in case they are different.

Having defined the problem definition formally, we still need to say what a solution to it is. Solutions to model reconciliation explanation problems are called *explanations*, which in turn are defined based on *model updates*, which we define next. A model update updates the *human's* model $\mathcal{M}_h^R$ to make it align with the actual model, i.e., the one of the robot, $\mathcal{M}^R$. Formally, such updates are defined as follows:

Definition 4. *For a given model reconciliation problem $\mathcal{P}^{MRE} = \langle \mathcal{M}^R, \mathcal{M}_h^R, \pi_R^* \rangle$, a **model update** is given by a tuple $\mathcal{E} = \langle \epsilon^+, \epsilon^- \rangle$, such that $\epsilon^+ \subseteq \Gamma(\mathcal{M}^R) \setminus \Gamma(\mathcal{M}_h^R)$ and $\epsilon^- \subseteq \Gamma(\mathcal{M}_h^R) \setminus \Gamma(\mathcal{M}^R)$. We will refer to the model $\mathcal{M}_h^R + \mathcal{E} = \Gamma^{-1}((\Gamma(\mathcal{M}_h^R) \setminus \epsilon^-) \cup \epsilon^+)$ as the updated human model that results from applying $\mathcal{E}$.*

We can now define solutions for model reconciliation explanation problems. Note that we don't require a "complete" set of changes to the human model making it *identical* to the actual one. It suffices to "explain" the robot's plan, i.e., so that this plan becomes optimal in the human's model.

Definition 5. *For a given model reconciliation explanation problem $\mathcal{P}^{MRE} = \langle \mathcal{M}^R, \mathcal{M}_h^R, \pi_R^* \rangle$, a model update $\mathcal{E} = \langle \epsilon^+, \epsilon^- \rangle$ is considered to be **a valid explanation** if the plan π_R^* is an optimal plan in $\mathcal{M}_h^R + \mathcal{E}$.*

2.4 Relevant Complexity Classes

While many of the standard results related to classical planning tend to either fall into NP or PSPACE classes [Bylander, 1994], the problem studied in this paper focuses on a class that is placed between these classes.

One way to view NP problems, is in terms of the existence of a witness or certificate that can be verified in polynomial time. Following Definition 2.1 by Arora and Barak [2009], a language L is said to be in NP if

$$x \in L \iff \exists u \in \{0,1\}^{p(|x|)} \text{ such that } M(x,u) = 1,$$

where M is a polynomial time Turing machine, p a polynomial and u is the witness. On the other hand, a language L is said to be in Σ_2^p [Arora and Barak, 2009, Definition 5.1] if

$$x \in L \iff \exists u \in \{0,1\}^{p(|x|)}\ \forall v \in \{0,1\}^{p(|x|)} \text{ such that } M(x,u,v) = 1,$$

where M is again a polynomial time Turing machine. Another way to view Σ_2^p is in terms of oracle machines.

The canonical Σ_2^p-complete problem is the special subclass of quantified boolean formula called *QSAT*$_2$ [Stockmeyer, 1976], where *QSAT*$_2$ corresponds to the question of satisfiability of a formula of the form

$$\exists X \forall Y \phi,$$

where X and Y are vectors over boolean variables and ϕ is a propositional formula defined over X and Y. Note that per Theorem 4.1 (1) from Stockmeyer [1976], *QSAT*$_2$ is complete for Σ_2^p regardless of the form. This fact is exploited in our membership proof as we map the $\mathcal{P}^{MRE}$ to a quantified boolean formula which is not necessarily in either CNF or DNF form. In fact, in the formula used for the membership proof, while $\phi_1(X)$ and $\phi_3(X,Z)$ are in CNF, the subformula $\neg\phi_3(X,Y)$ is a negation. Additionally, Theorem 4.1 (2) by Stockmeyer [1976] also shows that the subset *QSAT*$_2 \cap$ 3-DNF is also complete for Σ_2^p. So in our hardness proof rather than reducing arbitrary quantified boolean formulas into a $\mathcal{P}^{MRE}$, we focus on reducing an instance from the set *QSAT*$_2 \cap$ 3-DNF.

The polynomial hierarchy (PH) is formed by taking the union over the various classes of the form Σ_i^P, where each class Σ_i^P is defined in the above form with i alternating existential and universal quantifiers (starting with an existential quantifier as in the case of Σ_2^P). Finally, the PSPACE complexity class, covers all the problems that can be solved by a Turing machine with polynomial space [Arora and Barak, 2009]. A PSPACE-complete problem is the satisfiability of a TQBF or True Quantified Boolean Formula, where no restriction is placed on the quantification over the variables. While it is known that PH $\subseteq$ PSPACE holds, the problem of establishing PH $\neq$ PSPACE remains an important open problem (or question, as it's not known yet).

3 Complexity Results for Model Reconciliation Explanation Problems

We are interested in the computational complexity of solving model reconciliation explanation problems. It is however rather obvious that checking whether *any* solution exists is a trivial problem:

Proposition 1. *Let $\mathcal{P}^{MRE} = \langle \mathcal{M}^R, \mathcal{M}_h^R, \pi_R^* \rangle$ be a model reconciliation explanation problem. The question whether there exists a valid explanation can be decided in* constant time. *More precisely, the answer is always* yes.

The reason why the answer is always yes is because we could always simply compute the difference between the sets $\Gamma(\mathcal{M}_h^R)$ and $\Gamma(\mathcal{M}^R)$ and present these differences as explanation. (And we know that this is always possible, so we don't need to do so just to decide whether this explanation exists – it always does.) Thus, *computing* such an explanation is harder than deciding whether one exists; computing it is a linear problem. This corresponds to the class of explanation called *model patch explanation* previously studied in the literature [Chakraborti *et al.*, 2017].

While model patch explanations are technically correct, they might in practice not be the best explanations as it would involve resolving differences that are irrelevant, in that they didn't cause confusion. Recall that the reason for the necessity of explaining something is that the robot's plan either

isn't even a plan at all in the human's model, or just not optimal. So any explanation should restrict to finding reasons pointing to any of these facts.

Thus, what we are interested in is finding a *minimal* explanation, i.e., we want to present an explanation to the human user that involves the fewest possible number of model changes so that the observed plan is an optimal solution in his/her model – even if there are still some differences to the robot's model (of which the human would then still be unaware of). Such explanations have been referred to as *minimally complete explanations* [Chakraborti *et al.*, 2017].

To turn this *optimization problem* into a *decision problem*, we introduce (as it is usually done) an additional parameter representing the criterion that's being optimized – in our case the number of performed changes. Formally:

Definition 6. *For a given model reconciliation explanation problem $\mathcal{P}^{MRE} = \langle \mathcal{M}^R, \mathcal{M}_h^R, \pi_R^* \rangle$, we define the* ***optimal model reconciliation explanation decision problem*** *as:*

Given $\mathcal{P}^{MRE}$ and a natural number $k \in \mathbb{N} \cup \{0\}$, does there exist a valid explanation $\mathcal{E} = \langle \epsilon^+, \epsilon^- \rangle$ for $\mathcal{P}^{MRE}$, such that $|\epsilon^+| + |\epsilon^-| = k$? (We call this ***MRE-$k$****.)*

We are going to show that the problem is Σ_2^p-complete, which we show in the next two sections, one showing membership, the other showing hardness.

To prove the computational complexity, we will focus on the canonical Σ_2^p complete problem called *QSAT*$_2$ [Stockmeyer, 1976], where *QSAT*$_2$ corresponds to the question of satisfiability of a formula of the form $\exists X \forall Y \phi$, where X and Y are disjoint sets of boolean variables and ϕ is a propositional formula defined over X and Y. When we focus on specific forms propositional formula, say 3-DNF, we will denote it as *QSAT*$_2 \cap$ 3-DNF.

3.1 Membership Proof

Our first task would be to establish the fact that **MRE-k** is in a fact a member of the complexity class Σ_2^p. We will do so by reducing the problem into a *QSAT*$_2$ problem. In particular, by mapping a model reconciliation explanation problem $\mathcal{P}^{MRE} = \langle \mathcal{M}^R, \mathcal{M}_h^R, \pi_R^* \rangle$ into *QSAT*$_2$ of the form

$$\exists X, Z \phi_1(X) \wedge \neg(\exists Y \phi_2(X, Y)) \wedge \phi_3(X, Z),$$

such that $|X| = k \cdot |E^+ \cup E^-|$, where $E^+ = \Gamma(\mathcal{M}^R) \backslash \Gamma(\mathcal{M}_h^R)$ and $E^- = \Gamma(\mathcal{M}_h^R) \backslash \Gamma(\mathcal{M}^R)$ are the set of propositional variables that will capture the *possible* individual model updates, Y corresponds to the propositional variables required to encode a planning problem where the maximum plan length is limited to $|\pi_R^*| - 1$ that will be used to capture the possible shorter plans and Z includes the propositional variables required to encode a planning problem where the maximum plan length is limited to $|\pi_R^*|$ which will be used to encode the validity of π_R^*. Following the variables, $\phi_1(X)$ is a CNF formula that enforces the fact that only explanations of size k are possible, $\phi_2(X, Y)$ is a CNF formula that encodes whether given the explanation a plan of length $|\pi_R^*| - 1$ can achieve the goal in the updated model and finally, $\phi_3(X, Z)$ is a CNF formula that encodes whether given the explanation π_R^* is valid in the updated model. Applying the negation and moving the quantification upfront, we get the pre-nex *QSAT*$_2$ form

$$\exists X, Z \forall Y \phi_1(X) \wedge \neg \phi_2(X, Y) \wedge \phi_3(X, Z)$$

Now the important parts of this compilation are encoding the enforcement of explanation length (through $\phi_1(X)$) and encoding planning models in such a way that they can reflect the effects of model updates captured by a particular instantiation of X variables (used in $\phi_2(X, Y)$ and $\phi_3(X, Z)$.

Encoding Explanation Length

The variable set X consists of k propositional variables for each individual model update, i.e,

$$X = \bigcup_{\tau_i \in E^+ \cup E^-} \{\tau_i^1, ..., \tau_i^k\},$$

where each τ_i^m can be thought of as capturing the fact that the model update τ_i is applied at step m. Now our requirement is to enforce that only a single model update is applied at a given step. This is exactly done by $\phi_1(X)$, where $\phi_1(X)$ is a conjunction of clauses of the form

$$\tau_i^m \Rightarrow \bigwedge_{\tau_j \in E^+ \cup E^- \wedge \tau_j \neq \tau_1} \neg \tau_j^m.$$

$\phi_1(X)$ will contain a clause for every pair of model updates $\tau_i, \tau_j \in E^+ \cup E^-$ and every step $m \in \{1, ..., k\}$). Now $\phi_1(X)$ cannot be true if for any step more than one model update variable is true.

Encoding Planning Models Conditioned on Model Updates

Now the second part of the encoding requires that we have a way to capture the horizon-limited planning problem, that reflects the model updates captured by a given instantiation of X. This will be used in both $\phi_2(X, Y)$ and $\phi_3(X, Z)$. The encoding will be based on the planning-as-SAT encoding discussed Section 2.2. Let $\phi_T^{\mathcal{M}}$ be the unmodified original SAT encoding (in CNF form) corresponding to a model $\mathcal{M} = \langle F, A, \delta, I, G \rangle$ for a planning horizon T.

To allow for the model update, we will start with setting the encoding to be equal to the human model ($\phi_T^{\mathcal{P}^{MRE}} = \phi_T^{\mathcal{M}_h^R}$). We first augment the model component clauses in this model. Specifically, let ψ be a clause corresponding to a model component that is part of the human model but not part of the robot model, and let the corresponding model parameter be $\tau_j \in E^-$. Then we replace ψ in $\phi_T^{\mathcal{P}^{MRE}}$ with a clause

$$(\neg \tau_j^1 \wedge ... \wedge \neg \tau_j^k) \Rightarrow \psi,$$

This clause captures the fact that if the model component is not removed in the k explanation steps, then the model component should be considered when coming up with the plan. Note that this is still a clause as conjunction is on the left side of the implication.

Let ψ' be a clause corresponding to a model component that is part of the robot model but missing from the human model, with a corresponding model parameter be $\tau_j \in E^+$. We add a conjunction of the form given below to $\phi_T^{\mathcal{P}^{MRE}}$

$$(\tau_j^1 \Rightarrow \psi') \wedge ... \wedge (\tau_j^k \Rightarrow \psi')$$

That is, the model component needs to hold if the corresponding explanation is provided at any explanation step.

Now we will remove all the original explanatory frame axioms and add one that covers action definitions from both models, i.e., for ever fluent $f \in F^{\mathcal{M}^R}$ and each time step m up to $T-1$ we add a clause of the form $\neg f^m \wedge f^{m+1} \Rightarrow A^f_{add}$, where $A^f_{add} = \{a \mid a \in A^{\mathcal{M}^R} \wedge f \in add^{\mathcal{M}^R} \cup add^{\mathcal{M}^R_h}\}$. We can similarly add an explanatory fluent for the deletes.

Now to account for the explanation, we will add new clauses that will ensure that to use any previously missing adds or deletes at a time step, the respective model update should be performed at some explanation step or not performed at all if it was an effect that was not part of the robot model. Finally we leave the action exclusion clauses unmodified in the new model $\phi^{\mathcal{P}^{MRE}}_T$. This is sufficient as the human, robot and updated model all share the same action names. We can now see that this encoding is equivalent to the updated model.

Proposition 2. *Let $\vec{x}$ be a specific instantiation of the variable X corresponding to a model update $\mathcal{E} = \langle \epsilon^+, \epsilon^- \rangle$, such that $\epsilon^+ = \{\tau_1, ..., \tau_r\}$, $\epsilon^- = \{\bar{\tau}_1, ..., \bar{\tau}_p\}$ and $|\epsilon^+| + |\epsilon^-| = k$. Now let $\phi_{\vec{x}}(X)$ be a logical conjunction of the form*

$$x^1_{\tau_1} \wedge ... \wedge x^r_{\tau_r} \wedge x^{r+1}_{\bar{\tau}_1} \wedge ... \wedge x^k_{\bar{\tau}_p}$$

Now for the combined logical formula

$$\phi_{\vec{x}}(X) \wedge \phi_1(X) \wedge \phi^{\mathcal{P}^{MRE}}_T,$$

every instantiation of propositional variables that satisfies the formula corresponds to a plan for $\mathcal{M}^R_h + \mathcal{E}$.

This fact should be obvious from the validity of the original encoding. Any differences in the encoding are only those related to the explanations. For example, in the new encoding the enforcement of a positive precondition f for an action a_l at time step m that could be removed by a model update is going to be, $\neg\tau^1 \wedge ... \neg\tau^k \Rightarrow (a^m_l \Rightarrow f^{m-1})$, where $\tau = a_l$-*has-pos-prec-*f. So if τ^i is set true at any of the k sets then the precondition needs no longer to be satisfied.

We get $\phi_2(X, Y)$ by using the $\phi^{\mathcal{P}^{MRE}}_T$ encoding but for time horizon $|\pi^*_R| - 1$ (the encoding also includes NOOP actions to allow for shorter plans) and we define $\phi_3(X, Z)$ to be equal to

$$\phi^{\mathcal{P}^{MRE}}_{|\pi^*_R|}(X, Z) \wedge \phi_{\pi^*_R},$$

where $\phi_{\pi^*_R} = a^1_1 \wedge ... a^{|\pi^*_R|}_{|\pi^*_R|}$, a^i_i is the variable in Z corresponding to the i^{th} action in plan π^*_R for time step i.

Lemma 1. *A given problem $\mathcal{P}^{MRE}$ has a k-sized explanation if and only if*

$$\exists X, Z \forall Y \phi_1(X) \wedge \neg\phi_2(X, Y) \wedge \phi_3(X, Z)$$

is satisfiable.

This follows directly from the construction and Proposition 2. The formula is only satisfied if there exists an instantiation of X corresponding to an explanation of length k (enforced by $\phi_1(X)$) that allows for the validity of the current plan (enforced by $\phi_3(X, Z)$) and doesn't allow for any shorter plans (enforced by $\neg\phi_2(X, Y)$). This leads us to:

Theorem 1. ***MRE-**k is in Σ^p_2*

3.2 Hardness Proof

To prove that the problem **MRE-**k is Σ^p_2-hard, we will provide a polynomial reduction of a *QSAT*$_2 \cap$ 3-DNF instance (which has been shown to be Σ^p_2-complete [Stockmeyer, 1976]) into a k-bounded Model Reconciliation Explanation problem $\mathcal{P}^{MRE}$ thus solving **MRE-**k. To present the reduction, consider an arbitrary *QSAT*$_2 \cap$3-DNF instance $\exists X \forall Y\ \phi$, where ϕ is a disjunction consisting of N disjuncts (i.e., conjunctions) denoted as $C_1, .., C_N$ (each of size 3 as ϕ is in 3-DNF) defined over the propositions in X and Y.

As mentioned earlier, we will map the problem of checking the satisfiability of a propositional formula over a universal quantification to that of an optimality check. In particular, we will construct a planning model where the plan space covers the space of all possible instantiations of the universally quantified variables. We then establish the satisfiability of the universally quantified formula by showing that no plan (i.e., a specific instantiation of the variables) can satisfy the negation of the formula. This follows from the fact that

$$\forall Y \phi \iff \neg(\exists Y \neg\phi)$$

Though in our case, there is not only a universally quantified variable set Y but an existentially quantified variable set X. We will map this existentially quantified variable set into the initial state of the problem and will allow the model reconciliation problem to select any possible instantiation of X as part of the explanation. Thus mapping the problem to

$$\exists X \forall Y \phi \iff \exists X \neg(\exists Y \neg\phi)$$

That is, a valid explanation will show that there exists an initial state for which there exists no shorter action sequence that can satisfy the goal $\neg\phi$. We will also add some additional constraints in the two models that will form our model reconciliation explanation problem $\mathcal{P}^{MRE}_{QSAT}$ to ensure that the plan being explained will be optimal after all the differences between the models have been resolved.

The exact construction of the $\mathcal{P}^{MRE}_{QSAT}$ problem is given below. Let us first start by defining the fluent set and the action names. Let the fluent set F be defined as

$$F = F_X \cup F_Y \cup F_N \cup F_G \cup F_{D_1} \cup F_{D_2} \cup F_S,$$

where F_X and F_Y are the sets of fluents corresponding to X and Y, respectively (i.e., we could just set $F_X = X$ and $F_Y = Y$), F_N consists of a fluent per disjunct in ϕ (i.e., we could just set $F_N = \{C_1, \ldots, C_N\}$), $F_G = \{g\}$ contains a single goal fluent to be used by the models, F_{D_1} is a set of dummy fluents such that $|F_{D_1}| = |X| + 1$, F_{D_2} is a second set of dummy fluents such that $|F_{D_2}| = |Y| + N + 1$ and $F_S = \{p_s\}$ is a staging variable used to enforce action ordering. Now the action names A to be shared between the two models would be such that $|A| = |Y| + 3 \cdot |F_N| + |F_{D_2}| + |F_{D_1}| + 1$, where we would have an action for each of the fluents in the corresponding fluent subsets Y and F_{D_2}, and an action for each conjunction of the propositional formula ϕ. Finally, there are $|F_{D_1}| + 1$ "goal actions", where there are $|F_{D_1}|$ goal actions by which the goal can be established if $\neg\phi$ can be achieved and there is one goal action to be used as part of π. Specifically, we will represent the action names as follows

$$A = A_Y \cup A_{\neg\phi} \cup A_{(G,\neg\phi)} \cup A_{D_2} \cup A_{(G,D_2)}$$

Now we will define a model reconciliation explanation problem $\mathcal{P}^{MRE}_{QSAT} = \langle \mathcal{M}_1, \mathcal{M}_2, \pi \rangle$. The models are defined as $\mathcal{M}_1 = \langle F, A, \delta, I^{\mathcal{M}_1}, G \rangle$ and $\mathcal{M}_2 = \langle F, A, \delta, I^{\mathcal{M}_2}, G \rangle$, where $I^{\mathcal{M}_1} = X \cup F_S$ and $I^{\mathcal{M}_2} = F_{D_1} \cup F_S$. Note that the models only differ in their initial states.

We will now go on defining each of the actions. Starting with A_Y, an action $a^i_Y \in A_Y$ (for a variable $y_i \in Y$) is defined by $pre_+(a^i_Y) = F_S, pre_-(a^i_Y) = \emptyset, add(a^i_Y) = \{f^i_Y\}$, and $del(a^i_Y) = \emptyset$. That is, the action definition for a^i_Y is empty but for a single add effect that sets the fluent corresponding to the variable y_i true ($f^i_Y \in F_Y$) and a positive precondition that requires the staging variable to be true.

Next come the actions in $A_{\neg\phi}$, which will help us test whether for a given instantiation of X and Y, the negation of the propositional formula $\neg\phi(x, y)$ is satisfiable. Note that when we negate the 3-DNF ϕ, we obtain a 3-CNF $\neg\phi$, where each (i^{th}) conjunction C_i gets turned into a disjunction (clause) C'_i given by $\{p^i_1, p^i_2, p^i_3\}$ (where the literals got inverted, i.e., switched from negated to positive and vice versa). Now for each literal in C'_i we will define an action $a^{i,j}_{\neg\phi}$, $1 \leq j \leq 3$, as follows.

- if p^i_j is positive we have:
  $pre_+(a^{i,j}_{\neg\phi}) = \{f^j_{X\cup Y}\}, pre_-(a^{i,j}_{\neg\phi}) = \emptyset,$
  $add(a^{i,j}_{\neg\phi}) = \{f^i_N\}$, and $del(a^{i,j}_{\neg\phi}) = F_S$
- and if p^i_j is negative then we define it as
  $pre_+(a^{i,j}_{\neg\phi}) = \emptyset, pre_-(a^{i,j}_{\neg\phi}) = \{f^j_{X\cup Y}\},$
  $add(a^{i,j}_{\neg\phi}) = \{f^i_N\}$, and $del(a^{i,j}_{\neg\phi}) = F_S$,

where $f^i_N \in F_N$ is an indicator variable identifying whether the clause C'_i is satisfied and $f^j_{X\cup Y} \in F_X \cup F_Y$ is the fluent corresponding to the proposition. That is, the fluent becomes a positive precondition if it was a positive literal in the clause (resulting from negating the conjunction), otherwise it becomes a negative precondition. If at least one of the literals in the clause is satisfied the clause is satisfied (captured by the add effect). Additionally, the action deletes the staging variable p_s. This allows us to ensure that no action from A_Y can be performed after executing an action from the set $A_{\neg\phi}$.

Now one way for the goal g to be satisfied would be to satisfy all the negated clauses in ϕ, which is captured by the set of actions $A_{(G,\neg\phi)} = \{a^1_{(G,\neg\phi)}, ..., a^{|X|+1}_{(G,\neg\phi)}\}$. Here we have a possible goal action for each fluent in F_{D_1}. The actions here are defined such that $pre_+(a^i_{(G,\neg\phi)}) = F_N \cup \{f^i_{D_1}\}$, $pre_-(a^i_{(G,\neg\phi)}) = \emptyset$, $add(a^i_{(G,\neg\phi)}) = \{g\}$, and $del(a^i_{(G,\neg\phi)}) = \emptyset$, where $f^i_{D_1} \in F_{D_1}$. As we will see, once we complete the mapping of the satisfaction problem into the explanation problem, all the shorter action sequences that the explanation would need to invalidate would use these goal generating actions.

The optimal plan π that needs to be explained is given by $\pi = \langle a^1_{D_2}, ..., a^{|F_{D_2}|}_{D_2}, a_{(G,D_2)} \rangle$. This plan contains all the actions that are part of $A_{D_2} = \{a^1_{D_2}, ..., a^{|F_{D_2}|}_{D_2}\}$ and $A_{(G,D_2)} = \{a_{(G,D_2)}\}$, such that

- $pre_+(a^i_{D_2}) = pre_-(a^i_{D_2}) = \emptyset$, and for all $f^i_{D_2} \in F_{D_2}$: $add(a^i_{D_2}) = \{f^i_{D_2}\}$, $del(a^i_{D_2}) = \emptyset$, and
- $pre_+(a_{(G,D_2)}) = F_{D_2}$, $pre_-(a_{(G,D_2)}) = \emptyset$, $add(a_{(G,D_2)}) = \{g\}$, and $del(a_{(G,D_2)}) = \emptyset$.

This brings us to the important properties (captured by the following propositions and lemmata) that will let us establish the soundness of the reduction.

Proposition 3. *π is optimal in $\mathcal{M}_1$.*

In $\mathcal{M}_1$ none of the variables in F_{D_1} are true, neither can any action turn it true. Thus none of the goal actions in $A_{(G,\neg\phi)}$ can be used, leaving the model to use all actions in π to achieve the goal.

Proposition 4. *One can reach a state that captures (in terms of the truth values of F_Y) any possible instantiation of the variables Y by a plan of length less than $|\pi| - N - 1$.*

The proposition follows directly given the size of $|A_Y|$ and the fact that the actions in this set do not have preconditions.

Proposition 5. *Any possible instantiation of the variables X can be captured by the initial state of the updated model formed from a model update of size $|X|$.*

Note that the model update takes the form of $\langle \epsilon^+, \epsilon^- \rangle$, such that $\epsilon^+ \subseteq X$ and $\epsilon^- \subseteq F_{D_1}$. Thus one can create an explanation that sets some subset of $X' \subseteq X$ true, by making ϵ^+ equal to that subset (in terms of F_X) and selecting a subset of F_{D_1} as ϵ^-, such that $|\epsilon^-| = |X| - |\epsilon^+|$. Since $|F_{D_1}| = |X| + 1$ and $|\epsilon^-|$ is a non-negative number upper-bounded by $|X|$, we can always find a subset of F_{D_1} of size $|X| - |\epsilon^+|$.

Proposition 6. *For a model update $\mathcal{E} = \langle \epsilon^+, \epsilon^- \rangle$, where X^{ϵ^+} contains the values from X corresponding to the update, there exists an action sequence π' that is a valid plan in $\mathcal{M}_2 + \mathcal{E}$ such that $|\pi'| < |\pi|$, if and only if there exists some instantiation of variables Y (say $Y^{\pi'}$), such that for X^{ϵ^+} and $Y^{\pi'}$ it satisfies $\neg\phi$ (denoted as $(X^{\epsilon^+}, Y^{\pi'}) \models \neg\phi$).*

This follows directly from the fact that an action sequence of length less than $|\pi|$ can only satisfy the goal by using an action in $A^{\neg\phi}_G$, which requires satisfying all the clauses in $\neg\phi$. Similarly all instantiations of Y and testing validity of $\neg\phi$ can be done in less than $|\pi|$ steps, which brings us to the lemma:

Lemma 2. *For the model reconciliation explanation problem $\mathcal{P}^{MRE}_{QSAT} = \langle \mathcal{M}_1, \mathcal{M}_2, \pi \rangle$, there exists a valid explanation of size $|X|$ if and only if the corresponding QSAT$_2$ $\exists X \forall Y \phi$ is satisfiable.*

Proof. This lemma can be directly built from the previous three propositions. If there exists a valid explanation of size $|X|$ that means no plan of size less than $|\pi|$ is valid, that means there exists an instantiation of variables X (say $X^{\mathcal{E}}$), such that for all instantiations Y' of variables Y, we have

$$(X^{\mathcal{E}}, Y') \not\models \neg\phi$$

which means, we have for all Y' of Y

$$(X^{\mathcal{E}}, Y') \models \phi$$

Similarly if the QSAT$_2$ formula was not satisfiable, then for every $X^{\mathcal{E}}$ we should have at least one instantiation Y' for

which $(X^{\mathcal{E}}, Y') \models \neg\phi$. In the updated model we can now construct a plan that corresponds to Y' and the evaluation of $\neg\phi$ for Y' which should satisfy an action in $A_G^{\neg\phi}$. □

Which bring us to the theorem.

Theorem 2. ***MRE-**k is Σ_2^p-hard*

This theorem follows directly from Lemma 2 and the fact that *QSAT*$_2 \cap$ 3-DNF is Σ_2^p-complete. Finally, Theorem 1 and Theorem 2 brings us to our central result.

Theorem 3. ***MRE-**k is Σ_2^p-complete*

4 Related Work

While the complexity of the original model-reconciliation explanation has gone unexplored, there are complexity results from related problems that give us some clues about the actual complexity. For one, Sreedharan *et al.* [2020], showed PSPACE-completeness for providing a plan and a set of model updates such that the plan is valid in both the robot and the updated human model. Of course, here there is the additional complexity of identifying the plan and the problem overlooks the complexity of establishing the optimality, a central concern in the original model reconciliation formulation [Chakraborti *et al.*, 2017]. On the other hand, Lin and Bercher [2021] established that the complexity of updating the model (with any or a minimal number of changes) ensuring the validity of a given plan is an NP-complete problem (for most cases, only a few are in P). However, they do not constrain the model updates to those that align with the target robot model directly, as this is only implicitly provided via the plan that's supposed to be valid. Finally, Vasileiou *et al.* [2021], looked at a variant of model reconciliation that is framed as the problem of finding the shortest logical support of a given propositional formula in the context where the human and robot model are represented as propositional knowledge bases. They discuss a possible membership of this problem in the Σ_2^p class, but unfortunately, again the problem they define is different from the one studied by Chakraborti *et al.* [2017]. Vasileiou *et al.* [2021] looked at a case where there exists a knowledge base associated with the system KB_a and one associated with the human KB_h and there is a logical formula ϕ that needs to be explained such that $KB_a \models \phi$ and $KB_h \not\models \phi$. The explanation here takes the form of a support $\epsilon \subseteq KB_a \wedge KB_h$ such that $KB_h \wedge \epsilon \models \phi$ (as discussed by Vasileiou *et al.* [2021] in Definition 7). Given the generality of this formulation, one could map explanations for a classical planning problem partially into this framework, however it is not the exact problem studied by Chakraborti *et al.* [2017].

The first point to note is the fact that the explanation here doesn't support removal of rules from the human's knowledge base, a key form of model update discussed by Chakraborti *et al.* [2017]. While the definition does not allow for such a change, their algorithm (Algorithm 2) does include an ad-hoc test for satisfiability of $KB_a \wedge KB_h$ and the algorithm allows for the removal of formulas from KB_h if the conjunction is unsatisfiable. But this doesn't cover all the changes that are supported by Chakraborti *et al.* [2017]. For example, consider a case when an action a has an additional add effect e in the human knowledge base, which manifests in the form of two rules, a rule of the form

$$a^i \Rightarrow e^i$$

and a rule in the explanatory frame axiom that says a is a possible action that can satisfy the transition from $\neg e^*$ to e^*. In this case, the problem encoding of length n (n being the length of the plan being explained), the formula $KB_a \wedge KB_h$ needs not be unsatisfiable, especially if the plan being explained is not using a or e. However this rule could prevent some ϕ, say there doesn't exist a plan of length shorter than n, from being entailed without its removal. For example, let there be an action a_2 that directly sets the goal true if e is satisfied and there are no actions in KB_a that could have satisfied e. Now without removing this additional add effect there will always be a shorter plan. Unfortunately, after the first satisfiability test Vasileiou *et al.* [2021] don't provide any other way of removing rules from the human's knowledge base.

Next the paper breaks down the problem of explaining optimality of the plan into two separate problems. First it explains the validity and then it explains optimality. For explaining validity they mention adding the plan and the goal as constraints into the planning model as additional clauses and then testing for satisfiability. If the encoding is unsatisfiable, the authors mention that they "add the missing actions as part of the explanation" [Vasileiou *et al.*, 2021, page 5]. If this means they add all missing information in KB_h for actions in the plan, this will already result in more information that the one considered in the original model reconciliation work [Chakraborti *et al.*, 2017]. However, if they have some procedure to find the minimal information needed to be added to ensure optimality, this could still result in longer explanations. This is because choices made to ensure validity has an impact on the information needed to be provided to ensure optimality. As such the problem of finding model updates to ensure validity and optimality cannot be separated. For example, consider a case where the plan is invalid because a precondition for an action is unsatisfied in the human model. Now assume there are two possible minimal updates to ensure validity. One could say that the unsatisfied precondition is not part of that action in the knowledge base KB_h or there exists a previously missing effect of an earlier action that can satisfy this precondition (though it's not needed in KB_a). However one could build the rest of the model in such a way that adding the add effect information would result in other shorter plans being feasible. Say there are actions which can satisfy the goal directly whose only precondition in KB_h is satisfied by this effect, while those actions have extra preconditions in KB_a. This means the choice to introduce the add effect could make the explanation longer. Similarly we can create domains where the choice to remove the precondition would result in longer explanations.

5 Conclusion

One of the immediate points of interest is the comparison of the complexity of model-reconciliation with the complexity of classical planning. When no information about the problem to solve is given, worst case complexity is PSPACE-complete [Bylander, 1994]. Unless the polynomial hierarchy

collapses the problem of model reconciliation is thus easier than simple plan existence. This comparison is however not perfectly fair since in model reconciliation we have a plan given as an input thus bounding possible changes. While bounded plan existence is still PSPACE-complete since plan length can be bounded logarithmically, it turns NP-complete when encoded unarily (thus simulating the situation of model reconciliation where the bound is not given as number but explicitly via an input plan). Model reconciliation is thus harder than the respective bounded plan existence problem unless the polynomial hierarchy collapses.

Another consequence of the proof is the existence of an alternate problem formulation for generating a minimally complete explanation, namely by mapping it to $QSAT_2$ of increasing explanation length. This means one could use fast quantified boolean formula solvers to generate such explanations. One of the future directions for the work may be to investigate whether the use of this compilation provides an advantage over the A^* model space search proposed by Chakraborti *et al.* [2017]. Going forward it may also be worth investigating special cases of the model reconciliation explanation (restrictions on the planning model) that might be more tractable and consider the hardness of the various extensions of the model-reconciliation framework like those discussed by Sreedharan *et al.* [2018; 2019].

Another interesting direction of future work is generating lies as formalized by Chakraborti and Kambhampati [2019], which is also closely related to model reconciliation. In this case, the objective of the model updates remains the same, i.e., the robot is trying to ensure the plan is optimal in the updated model, but the model-updates is no longer required to be part of the robot's true model.

Acknowledgments

This research is supported in part by ONR grants N00014- 16-1-2892, N00014-18-1- 2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF19-2-0006 and a JP Morgan AI Faculty Research grant.

References

[Arora and Barak, 2009] Sanjeev Arora and Boaz Barak. *Computational Complexity - A Modern Approach*. Cambridge University Press, 2009.

[Bylander, 1994] Tom Bylander. The computational complexity of propositional STRIPS planning. *Artif. Intell.*, 69(1-2):165–204, 1994.

[Chakraborti and Kambhampati, 2019] Tathagata Chakraborti and Subbarao Kambhampati. (when) can ai bots lie? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 53–59, 2019.

[Chakraborti *et al.*, 2017] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI 2017*, pages 156–163. IJCAI Organization, 2017.

[Chakraborti *et al.*, 2020] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The emerging landscape of explainable automated planning & decision making. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4803–4811. IJCAI Organization, 2020.

[Fox *et al.*, 2017] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable Planning. In *IJCAI XAI Workshop*, pages 24–30. XAI, 2017.

[Geffner and Bonet, 2013] Hector Geffner and Blai Bonet. *A concise introduction to models and methods for automated planning*, volume 7 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2013.

[Hoffmann and Magazzeni, 2019] Jörg Hoffmann and Daniele Magazzeni. Explainable AI planning (XAIP): overview and the case of contrastive explanation (extended abstract). In *Reasoning Web. Explainable Artificial Intelligence*, volume 11810 of *Lecture Notes in Computer Science*, pages 277–282. Springer, 2019.

[Kautz *et al.*, 1996] Henry A. Kautz, David A. McAllester, and Bart Selman. Encoding plans in propositional logic. In *KR 96*, pages 374–384. Morgan Kaufmann, 1996.

[Lin and Bercher, 2021] Songtuan Lin and Pascal Bercher. Change the world – how hard can that be? on the computational complexity of fixing planning models. In *IJCAI 2021*, pages 4152–4159. IJCAI Organization, 2021.

[Sreedharan *et al.*, 2018] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Handling model uncertainty and multiplicity in explanations via model reconciliation. In *ICAPS 2018*, pages 518–526. AAAI Press, 2018.

[Sreedharan *et al.*, 2019] Sarath Sreedharan, Alberto Olmo Hernandez, Aditya Prasad Mishra, and Subbarao Kambhampati. Model-free model reconciliation. In *IJCAI 2019*, pages 587–594. IJCAI Organization, 2019.

[Sreedharan *et al.*, 2020] Sarath Sreedharan, Tathagata Chakraborti, Christian Muise, and Subbarao Kambhampati. Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning. In *AAAI 2020*, pages 2518–2526. AAAI Press, 2020.

[Sreedharan *et al.*, 2021] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301:103558, 2021.

[Stockmeyer, 1976] Larry J Stockmeyer. The polynomial-time hierarchy. *Theoretical Computer Science*, 3(1):1–22, 1976.

[Vasileiou *et al.*, 2021] Stylianos Loukas Vasileiou, Alessandro Previti, and William Yeoh. On exploiting hitting sets for model reconciliation. In *AAAI 2021*, pages 6514–6521. AAAI Press, 2021.