On the Relationship Between KR Approaches for Explainable Planning

Abstract

In this paper, we build upon notions from *knowledge representation and reasoning* (KR) to expand a preliminary logic-based framework that characterizes the model reconciliation problem for explainable planning. We also provide a detailed exposition on the relationship between similar KR techniques, such as abductive explanations and belief change, and their applicability to explainable planning. Finally, we provide preliminary experimental results that demonstrate the promise of our approach on problems that require long explanations.

Introduction

As there is a substantial need for transparency and trust between intelligent systems and humans, *Explainable AI Planning* (XAIP) has recently gained a lot of attention due to its potential adoption in real-world applications. Within this context, a popular theme that has recently emerged is called *model reconciliation* (Chakraborti *et al.* 2017). Researchers in this area have looked at how an agent can explain its decisions to a human user who might have a different understanding of the same planning problem. These explanations bring the model of the human user closer to the agent's model by transferring a minimum number of updates from the agent's model to the human's model. However, a common thread across most of these works is that they, not surprisingly, employ mostly automated planning approaches.

In this work, we tackle the *model reconciliation problem* (MRP) from the perspective of *knowledge representation and reasoning* (KR), where we extend a logic-based framework proposed by Vasileiou *et al.* (2019) and argue that it can effectively model the MRP. As our framework builds upon various KR techniques, we further give a detailed exposition on the relationship between such techniques and their applicability to XAIP as well as provide two examples that highlight the differences with our proposed framework.

Background

Logic: A *logic* L is a tuple $\langle KB_L, BS_L, ACC_L \rangle$, where KB_L is the set of well-formed knowledge bases (or theories) of L – each being a set of formulae. BS_L is the set of possible belief sets; each element of BS_L is a set of syntactic elements representing the beliefs L may adopt. $ACC_L : KB_L \rightarrow 2^{BS_L}$ describes the *"semantics"* of L by assigning to each element of KB_L a set of acceptable sets of beliefs. For each $KB \in KB_L$ and $B \in ACC_L(KB)$, we say that B is a *model* of KB. A logic is monotonic if $KB \subseteq KB'$ implies $ACC_L(KB') \subseteq ACC_L(KB)$.

Definition 1 (Skeptical Entailment) *A formula φ in the logic L is* skeptically entailed *by KB, denoted by $KB \models^s_L \varphi$, if $ACC_L(KB) \neq \emptyset$ and $\varphi \in B$ for* every $B \in ACC_L(KB)$.

Definition 2 (Credulous Entailment) *A formula φ in the logic L is* credulously entailed *by KB, denoted by $KB \models^c_L \varphi$, if $ACC_L(KB) \neq \emptyset$ and $\varphi \in B$ for* some $B \in ACC_L(KB)$

Definition 3 (Consistent Knowledge Base) *A KB is* consistent *iff $ACC_L(KB) \neq \emptyset$ or, equivalently, iff KB does not skeptically entail false.*

Throughout this paper, we consider a finitary propositional language L and represent a knowledge base by a propositional formula KB. For our later use, we will assume that a negation operator $\neg$ over formulas exists. Additionally, φ and $\neg\varphi$ are contradictory with each other in the sense that, for any KB and $B \in ACC_L(KB)$, if $\varphi \in B$, then $\neg\varphi \notin B$; and if $\neg\varphi \in B$, then $\varphi \notin B$. Therefore, if $\{\varphi, \neg\varphi\} \subseteq KB$, then KB is inconsistent, i.e., $ACC_L(KB) = \emptyset$. $\epsilon \subseteq KB$ is called a *sub-theory* of KB. A theory KB *subsumes* a theory KB', denoted by $KB \lhd KB'$, if $ACC_L(KB) \subset ACC_L(KB')$.

Classical Planning as Boolean Satisfiability: A classical planning problem can be naturally encoded as an instance of propositional satisfiability (Kautz and Selman 1992; Kautz *et al.* 1996). The basic idea is the following: Given a planning problem P, find a solution for P of length n by creating a propositional formula that represents the initial state, goal state, and the action dynamics axioms for n time steps. This is referred to as the *bounded planning problem* (P, n), and we define the formula for (P, n) such that: *Any* model of the formula represents a solution to (P, n) and if (P, n) has a solution, then the formula is satisfiable.

Finally, we can *extract* a plan by finding an assignment of truth values that satisfies Φ (i.e., for $i = 0, \ldots, n-1$, there is exactly one action a such that $a_i = \textit{True}$). Due to space constraints, we refer the reader to Kautz *et al.* (1996) for more details.

Model Reconciliation Problem (MRP): An MRP is defined by the tuple $\Psi = \langle \pi, \Phi \rangle$, where π is the optimal plan in M^R and $\Phi = \langle M^R, M_H^R \rangle$ is a tuple of the agent's model $M^R = \langle D^R, I^R, G^R \rangle$ and the agent's approximation of the human's model $M_H^R = \langle D_H^R, I_H^R, G_H^R \rangle$. A solution to MRP is an explanation ϵ such that when it is used to update the human's model M_H^R to $\widehat{M}_H^{R,\epsilon}$, the plan π is optimal in both the agent's model M^R and the updated human model $\widehat{M}_H^{R,\epsilon}$. The goal is to find a shortest explanation.

Logic-based Explanations in Planning

We now describe our framework, which generalizes the preliminary framework proposed by Vasileiou *et al.* (2019), that solves the model reconciliation problem by computing cost-minimal explanations with respect to two knowledge bases. We formulate the notion of explanation in the following setting, where, for brevity, we use the term $\models_L^x$ for $x \in \{s, c\}$ to refer to skeptical (s) or credulous (c) entailment:

> **Explanation Generation Problem:** Given two knowledge bases KB_a and KB_h and a formula φ in a logic L. Assume that $KB_a \models_L^x \varphi$ and $KB_h \not\models_L^x \varphi$. The goal is to identify an explanation (i.e., a set of formulas) $\epsilon \subseteq KB_a$ such that when it is used to *update* KB_h to $\widehat{KB}_h^{\epsilon}$, the updated $\widehat{KB}_h^{\epsilon} \models_L^x \varphi$.

When updating a knowledge base KB with an explanation ϵ, the updated knowledge base $KB \cup \epsilon$ may be inconsistent as there may be contradictory formulas in KB and ϵ. As such, to make KB consistent again, one needs to remove this set of contradictory formulae $\gamma \subseteq KB$ from KB. Further, it is vital that the KB must contain all formulae pertaining to the action dynamics of the given planning problem (see (Kautz *et al.* 1996)).

Definition 4 (Knowledge Base Update) *Given a knowledge base KB and an explanation ϵ, the updated knowledge base is $\widehat{KB}^{\epsilon} = KB \cup \epsilon \setminus \gamma$, where $\gamma \subseteq KB$ is the minimal set of formulae that must be removed from KB such that the updated $\widehat{KB}^{\epsilon}$ is consistent and contains all axioms that characterize the given planning problem (Kautz* et al. *1996).*

We now define the notion of a *support* of a formula w.r.t. a knowledge base before defining the notion of *explanations*.

Definition 5 (Support) *Given a knowledge base KB and a formula φ in logic L, assume that $KB \models_L^x \varphi$. We say that $\epsilon \subseteq KB$ is an* x-support *of φ w.r.t. KB if $\epsilon \models_L^x \varphi$. Assume that ϵ is an x-support of φ w.r.t. KB. We say that $\epsilon \subseteq KB$ is a* $\subseteq$-minimal x-support *of φ if no proper sub-theory of ϵ is an x-support of φ. Furthermore, ϵ is a* $\lhd$-general x-support *of φ if there is no support ϵ' of φ w.r.t. KB such that ϵ subsumes ϵ'.*

Definition 6 (Explanation) *Given two knowledge bases KB_a and KB_h and a formula φ in logic L, assume that $KB_a \models_L^x \varphi$ and $KB_h \not\models_L^x \varphi$. An* explanation for φ from KB_a for KB_h *is a support ϵ w.r.t. KB_a for φ such that the updated $\widehat{KB}_h^{\epsilon} \models_L^x \varphi$, where $\widehat{KB}_h^{\epsilon}$ is updated according to Definition 4.*

Explanations in Planning

Even though explanations can be composed for arbitrary queries, in this paper, we identify two important problems: (1) Explaining the *validity* of a plan to the human user, and (2) Explaining the *optimality* of a plan to the user. Naturally, the former problem must be solved before the latter problem since the user must accept that the plan is valid before they accept that the plan is optimal. However, it may be the case that only the former problem must be solved, especially when plan optimality is not a major concern to the user. From now until the end of this section, we use KB_a and KB_h to denote the knowledge bases encoding the planning problem of the planning agent and the human user, respectively.

Plan Validity: The question of plan validity can be formulated as follows. Assume $\pi = \{\alpha_1, \ldots, \alpha_n\}$ such that $KB_a \models_L^c \pi$ and $KB_h \not\models_L^c \pi$ (i.e., KB_h has no knowledge about the plan π). This lets us define plan validity below.

Definition 7 (Plan Validity) *Let Π be a planning problem, π a plan of Π, and KB_h a knowledge base encoding Π. If $KB_h \models_L^c \pi$, then we say that π is a valid plan in KB_h.*

Plan Optimality: Assume that π^* is an optimal plan in a model of KB_a. To explain the optimality of π^* to KB_h, we need to prove that no shorter plan exists in KB_h. Thus, we need to prove that no shorter plan exists in *all* models of KB_h. This can be done by utilizing skeptical entailment.

Definition 8 (Plan Optimality) *Let Π be a planning problem, π^* an optimal plan of Π with length n, and KB_h a knowledge base encoding Π. We say that π^* is optimal in KB_h if and only if $KB_h \models_L^c \pi^*$ and $KB_h \models_L^s \bigwedge_{t=0}^{n-1} \neg g_t$, where g_i is the fact corresponding to the goal of the planning problem at time step i.*

A Simple Approach for Restoring Consistency: When updating KB_h using Definition 6, it might be necessary to retract some formulae from KB_h to guarantee consistency. To efficiently solve this problem, we exploit a simple observation: There exists only a single model in a knowledge base encoding a planning problem that is consistent with an optimal plan π^* for that problem. The reason is that all facts are initialized by the start state and cannot change between subsequent time steps unless there are effects of actions that are executed. Further, only one action can be taken at each time step and all actions are deterministic. Using this observation, we generalize that the formulae in KB_h that are false according to this model must be erroneous with respect to KB_a. A trivial approach would be to use that model to identify the erroneous formulae and replace them with the corresponding (correct) formulae from KB_a. Thus, specifying a cost-function that minimizes the complexity of an explanation, i.e., w.r.t. subset-cardinality, this framework can be used to model the MRP and yield minimal explanations.

Optimal Plan Length	Explanation Length $\|\epsilon\|$									
	2		4		6		8		10	
	CSZK	ALG	CSZK	ALG	CSZK	ALG	CSZK	ALG	CSZK	ALG
6	**0.5s**	1.0s	2.0s	**0.9s**	9.5s	**0.8s**	300.0s	**1.0s**	500s	**1.0s**
10	**0.5s**	3.0s	**2.5s**	**2.5s**	9.5s	**3.0s**	300.0s	**3.5s**	500s	**2.5s**
12	**0.5s**	4.5s	**2.0s**	5.0s	9.0s	**6.5s**	305.0s	**4.5s**	505s	**7.0s**
16	**1.0s**	28.0s	**2.0s**	27.0s	**10.0s**	26.0s	309.0s	**27.0s**	502s	**31.0s**

Table 1: Varying Explanation and Plan Lengths for the BLOCKSWORLD PDDL Domain.

Key Differences with the Previous Framework: The key differences with the preliminary framework proposed by Vasileiou *et al.* (2019) are the following: (1) The previous framework was restricted to skeptical entailment while our generalized version applies to credulous entailment as well. (2) The previous framework assumes that $KB_h \subseteq KB_a$, which negates the need to remove any formula from KB_h during the update process to maintain consistency. In contrast, our framework makes no such assumption.

Preliminary Experimental Results

We now provide some preliminary experimental results of our approach, labeled ALG, against the current state-of-the-art algorithm by Chakraborti *et al.* (2017), labeled CSZK,[1] for solving model reconciliation problems. We imposed a timeout of 1800s. In our experiment, we varied the explanation length $|\epsilon|$ as well as the optimal plan length in the BLOCKSWORLD domain. Table 1 tabulates the results. We observe that the runtimes of CSZK increases as the explanation lengths increase. In contrast, the runtimes of ALG remain relatively unchanged with varying explanation lengths because they are dominated by the size of the encoded knowledge bases, which are independent of the explanation lengths. These preliminary results thus demonstrate the potential of our approach, especially on problems that require long explanations.

Relationship to Other KR Work

As our proposed framework bears some similarities with the theory of belief change and abductive explanations, in this section, we first describe their underlying theory. We then provide in the next section two examples that illustrate the differences between these approaches.

Abductive Explanations

Explanations in knowledge base systems were first introduced by Levesque (1989) in terms of abductive reasoning, that is, given a knowledge base and a formula that we do not believe at all, what would it take for us to believe that formula? A more formal definition is provided below.

Definition 9 (Abductive Explanation) *Given a knowledge base KB and a query q to be explained, α is an explanation of q w.r.t. to KB iff $KB \cup \{\alpha\}$ is consistent and $KB \cup \{\alpha\} \models^s_L q$.*

Usually, such explanations are phrased in terms of a hypothesis set H (set of atomic sentences – also called abducibles), and, generally, is an intuitive methodology for deriving root causes.

Belief Change

Belief change is a kind of change that can occur in a knowledge base. Depending on how beliefs are represented and what kinds of inputs are accepted, different typologies of belief changes are possible. In the most common case, when beliefs are represented by logical formulae, one can distinguish three main kinds of belief changes, namely, *expansion*, *revision*, and *contraction*. In the following, we formally describe the aforementioned notions.

Expansion: An expansion operator of a knowledge base can be formulated in a logical and set-theoretic notation:

Definition 10 (Expansion Operator) *Given a knowledge base KB and a formula ϕ, $+_e$ is an expansion operator if it expands KB by ϕ as $KB +_e \phi := \{\psi : KB \cup \phi \vdash \psi\}$.*

It is trivial to see that $KB +_e \phi$ will be consistent when ϕ is consistent with KB, and that $KB +_e \phi$ will be closed under logical consequences.

Revision: A belief revision occurs when we want to add new information into a knowledge base in such a way that, if the new information is inconsistent with the knowledge base, then the resulting knowledge base is a new consistent knowledge base. Alchourrón, Gärdenfors, and Makinson conducted foundational work on knowledge base revision, where they proposed a set of rationality postulates, called *AGM postulates*, and argued that every revision operator must satisfy them (Alchourrón and Makinson 1985; Gärdenfors 1986; Gärdenfors *et al.* 1995). Although revision cannot be defined in a set-theoretic manner closed under logical consequences (like expansion), it can be defined:

Definition 11 (Revision Operator) *Given a knowledge base KB and a formula ϕ, $+_r$ is a revision operator if it satisfies the AGM postulates for revision and modifies KB w.r.t. ϕ such that the resulting KB is consistent.*

The underlying motivation behind the AGM postulates is that when we change our beliefs, we want to retain as much as possible the information from the old beliefs. Thus, when incorporating new information in the knowledge base, the heuristic criterion should be the criterion of *information economy* (i.e., minimal changes to the knowledge base is preferred). As such, a model-theoretic characterization of minimal change has been introduced by Katsuno and Mendelzon (1991b), where minimality is defined as selecting the models of ϕ that are "closest" to the models of KB.

However, the AGM rationality postulates will not be adequate for every application. Katsuno and Mendelzon (1991a) proposed a new type of belief revision called *update*. The fundamental distinction between the two kinds of belief revision in a knowledge base, namely *revision* and *update*, is that the former consists of incorporating information about a static world, while the latter consists of inserting information to the knowledge base when the world described by it changes. As such, they claim that the *AGM postulates describe only revision* and showed that *update can be characterized by a different set of postulates called KM postulates*.

[1] We used the implementation of the authors, available at: https://github.com/TathagataChakraborti/mmp.

Definition 12 (Update Operator) *Given a knowledge base KB and a formula ϕ, $+_u$ is an update operator if it satisfies the KM postulates for update and modifies KB w.r.t. ϕ such that the updated KB incorporates the change in the world introduced by ϕ.*

From a model-theoretic view, the difference between revision and update, although marginal at first glance, can be described as follows: Procedures for revising KB by ϕ are those that satisfy the AGM postulates and select the models of ϕ that are "closest" to the models of KB. In contrast, update methods are exactly those that satisfy the KM postulates and select, for each model I of KB, the set of models of ϕ that are closest to I. Then, the updated KB will be characterized by the union of these models.[2]

It is worth mentioning that, on a high level, the key difference between update and revision is a temporal one: Update incorporates into the knowledge base the fact that the world described by it has changed, while revision is a change to our world description of a world that has not itself changed. We refer the reader to Katsuno and Mendelzon (1991a) for a comprehensive description as well as an intuitive meaning between revision and update.

Contraction: Similarly to the AGM postulates for revision, Alchourrón and Makinson (1985) proposed a set of axioms that any contraction operator must satisfy. Therefore, a contraction operator is defined by:

Definition 13 (Contraction Operator) *Given a knowledge base KB and a formula ϕ, $-_c$ is a contraction operator if it satisfies the AGM postulates for contraction, and contracts KB w.r.t. ϕ by retracting formulae in KB without adding of new ones.*

Interestingly, it has been shown that the problems of revision and contraction are closely related (Gärdenfors 1988). Despite the fact that the postulates that characterize revision and contraction are "independent,"[3] revision can be defined in terms of contraction (and vice versa). This is referred to as the Levi identity (Levi 1978).

Two Illustrative Examples

To illustrate the differences between our approach and the KR approaches described in the previous section, we discuss below how they operate on two planning examples.

Problem 1

Assume a planning problem with only one action $A = \{$precondition: P, effect: $E\}$ with initial and goal states P and E, respectively. Clearly, the plan for this problem is $\pi^* = [A]$. Also, assume that the human user is not aware that action A has effect E. Now, the knowledge bases encoding the models of the agent and the human, in the fashion of Kautz *et al.* (1996), are:

- $KB_a = [P_0, \neg E_0, E1, A_0 \rightarrow P_0, A_0 \rightarrow E_1, \neg E_0 \wedge E_1 \rightarrow A_0]$,
- $KB_h = [P_0, \neg E_0, E1, A_0 \rightarrow P_0]$.

Further, without loss of generality, suppose that the explanation needed to explain π^* to KB_h is $\epsilon = [A_0 \rightarrow E_1, \neg E_0 \wedge E_1 \rightarrow A_0]$.

Abductive Explanations: Abductive explanation cannot be applied in this setting because KB_h does not contain any causal rules that can be used to abduce the query.

Revision: Since the union of ϵ and KB_h is consistent, the revision operator will yield a trivial update according to the second AGM axiom: $KB_h +_r \epsilon = KB_h \cup \epsilon$.

Update: To use the update operator, we first need to find the models of KB_h and ϵ:

Models(KB_h): $I_1 = \{P_0, E_1, A_0\}$, $I_2 = \{P_0, E_1\}$.
Models(ϵ): $J_1 = \{A_0, E_1, P_0\}$, $J_2 = \{A_0, E_1\}$,
$J_3 = \{A_0, E_1, E_0, P_0\}$, $J_4 = \{A_0, E_1, E_0\}$,
$J_5 = \{E_1, E_0, P_0\}$, $J_6 = \{E_1, E_0\}$,
$J_7 = \{E_0, P_0\}$, $J_8 = \{E_0\}$, $J_9 = \{P_0\}$, $J_{10} = \{\}$.

Now, according to the KM postulates, we need to find the models of ϵ that are closest to I_1 and I_2. Then, the updated KB will be the disjunction of the conjunction of the variables in each model. Now, let the function *Diff*(m_1, m_2) denote the set of propositional letters with different truth values in models m_1 and m_2.

Then, for I_1, it is easy to see that the closest model is J_1 because *Diff*$(I_1, J_1) = \emptyset <$ *Diff*(I_1, J_k) for all k. So, J_1 is selected. For I_2, we need to calculate the difference for every model of ϵ:

$\underline{\textit{Diff}(I_2, J_1) = \{A_0\}}$, *Diff*$(I_2, J_2) = \{A_0, P_0\}$,
$\underline{\textit{Diff}(I_2, J_3) = \{A_0, E_0\}}$, *Diff*$(I_2, J_4) = \{A_0, E_0, P_0\}$,
$\underline{\textit{Diff}(I_2, J_5) = \{E_0\}}$, *Diff*$(I_2, J_6) = \{P_0, E_0\}$,
$\underline{\textit{Diff}(I_2, J_7) = \{E_0, E_1\}}$, *Diff*$(I_2, J_8) = \{E_0, E_1, P_0\}$,
$\underline{\textit{Diff}(I_2, J_9) = \{E_1\}}$, *Diff*$(I_2, J_{10}) = \{P_0, E_1\}$,

where sets with the minimal elements are underlined. So, J_1, J_5, and J_9 are selected and the final result is the union of all selected models, that is, *Models*$(KB_h +_u \epsilon) = \{J_1, J_5, J_9\}$. Thus, the resulting KB must satisfy all three models, yielding the following: $KB_h +_u \epsilon = [(A_0 \wedge E_1 \wedge P_0 \wedge \neg E_0) \vee (E_1 \wedge E_0 \wedge P_0 \wedge \neg A_0) \vee (P_0 \wedge \neg E_1 \wedge \neg A_0 \wedge \neg E_0)]$.

Our Approach: As a first step, our method will first check if KB_h is consistent with the model of KB_a. Since it is, it will simply insert ϵ to KB_h, yielding $\widehat{KB}_h^{\epsilon} = KB_h \cup \epsilon$ just like revision.

In conclusion, this problem demonstrates that it is possible for *belief revision* to yield the same update as our approach, which is when $KB_h \cup \epsilon$ is consistent (per AGM postulates). It also highlights why *belief update* is not applicable for explainable planning, namely that the updated knowledge base $KB_h +_u \epsilon$ violates the action dynamics of planning problems (Kautz *et al.* 1996).

Problem 2

Now assume a planning problem with the two actions $A = \{$precondition: P, effect: $G\}$ and $B = \{$precondition: E,

[2] This approach is called the *possible models approach (pma) (Winslett 1988).*

[3] In the sense that the postulates for revision do not refer to contraction and vice versa.

effect: $G\}$ with initial and goal states P and G, respectively, and a plan $\pi^* = [A]$. Also, assume that the human user is not aware that action A has effect G. Then, the knowledge bases encoding the models of the agent and the human are:

- $KB_a = [P_0, \neg E_0, \neg G_0, G_1, A_0 \rightarrow P_0, A_0 \rightarrow G_1, B_0 \rightarrow E_0, B_0 \rightarrow G_1, \neg G_0 \wedge G_1 \rightarrow A_0 \vee B_0, \neg A_0 \vee \neg B_0]$,
- $KB_h = [P_0, \neg E_0, \neg G_0, G_1, A_0 \rightarrow P_0, B_0 \rightarrow E_0, B_0 \rightarrow G_1, \neg G_0 \wedge G_1 \rightarrow B_0, \neg A_0 \vee \neg B_0]$.

As in the previous problem, we now assume that the explanation needed is $\epsilon = [A_0 \rightarrow G_1, \neg G_0 \wedge G_1 \rightarrow A_0 \vee B_0]$.

Abductive Explanations: The method of abductive explanations will fail in this setting due to the fact that KB_h is inconsistent. Further, even if it is consistent, we will still not be able to find any abductive explanations due to the lack of causal rules in KB_h.

Revision: Following AGM postulates, revision cannot be applied because KB_h is individually inconsistent.

Update: Again, as KB_h is inconsistent, and according to KM update postulates, it cannot be repaired using update.

Our Approach: As $KB_h \cup \epsilon$ is inconsistent, our approach will identify the erroneous formula $\neg G_0 \wedge G_1 \rightarrow B_0$ and replace it with the corresponding correct formula $\neg G_0 \wedge G_1 \rightarrow A_0 \vee B_0$ from KB_a, thereby restoring consistency. The updated knowledge base will be $\widehat{KB}_h^{\epsilon} = [P_0, \neg E_0, \neg G_0, G_1, A_0 \rightarrow P_0, A_0 \rightarrow G_1, B_0 \rightarrow E_0, B_0 \rightarrow G_1, \neg G_0 \wedge G_1 \rightarrow A_0 \vee B_0, \neg A_0 \vee \neg B_0]$.

In summary, this problem demonstrates that when KB_h is inconsistent, abductive explanations, revision, and update cannot be applied but our approach can be applied.

Discussion and Conclusions

A key distinction between the previous approaches and our approach is that, historically, belief change refers to a *single agent* revising its belief after receiving a new piece of information that is in conflict with its current beliefs; so, there is a temporal dimension in belief change and a requirement that it should maintain as much as possible the belief of the agent, per AGM postulates. Our notion of explanation is done with respect to *two knowledge bases* and there is no such requirement (with respect to KB_h). For example, if the agent believes that block A is on block B, the human believes that block B is on block A, and the explanation does not remove this fact from the human's KB, then the agent and the human will still have some conflicting knowledge about the positions of blocks A and B after the update. Thus, the previous notions of belief change cannot accurately capture and characterize the MRP problem.

Similar to belief change, explanation differs from other similar notions, such as diagnosis (Reiter 1987). In general, a diagnosis is defined with respect to a knowledge base KB, a set of components H, and a set of observations O. Given that $KB \cup O \cup \{\neg ab(c) \mid c \in H\}$ is inconsistent, a diagnosis is a subset S of H such that $KB \cup O \cup \{ab(c) \mid c \in S\} \cup \{\neg ab(c) \mid c \in H \setminus S\}$ is consistent. Here, $ab(c)$ denotes that the component c is faulty.

Generalizing this view, the inconsistency condition could be interpreted as the query q and $KB \cup O \models_L^s \neg q$. Then a diagnosis is a set $S \subseteq H$ such that $KB \cup O \cup S \models_L^s q$. An explanation for q from KB_a to KB_h is, on the other hand, a pair (S_1, S_2) such that $(KB_h \setminus S_2) \cup S_1 \models_L^s q$. Thus, the key difference is that an explanation might require the removal of some knowledge of KB_h while a diagnosis does not.

To conclude, we build upon notions from KR to expand a preliminary logic-based framework that characterizes the model reconciliation problem for explainable planning. We further provide a detailed exposition on the relationship between our framework and other similar KR techniques, such as abductive explanations and belief change, using two illustrative planning examples to describe their differences. Preliminary experimental results show that our framework can be faster than the current state-of-the-art solver on problems that require long explanations.

References

C. E. Alchourrón and D. Makinson. On the logic of theory change: Safe contraction. *Studia logica*, 1985.

T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI*, 2017.

P. Gärdenfors, H. Rott, DM Gabbay, CJ Hogger, and JA Robinson. Belief revision. *Computational Complexity*, 1995.

P. Gärdenfors. Belief revisions and the ramsey test for conditionals. *The Philosophical Review*, 1986.

P. Gärdenfors. *Knowledge in flux: Modeling the dynamics of epistemic states.* The MIT press, 1988.

H. Katsuno and A. O. Mendelzon. On the difference between updating a knowledge base and revising it. In *KR*, 1991.

H. Katsuno and A. O Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 1991.

H. Kautz and B. Selman. Planning as satisfiability. In *ECAI*, 1992.

H. Kautz and B. Selman. Pushing the envelope: Planning, propositional logic, and stochastic search. In *AAAI*, 1996.

H. Kautz, D. McAllester, and B. Selman. Encoding plans in propositional logic. In *KR*, 1996.

H. J. Levesque. A knowledge-level account of abduction. In *IJCAI*, 1989.

I. Levi. Subjunctives, dispositions and chances. In *Synthese*. 1978.

R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 1987.

S. L. Vasileiou, W. Yeoh, and S. C. Tran. A preliminary logic-based approach for explanation generation. In *2nd ICAPS XAIP Workshop*, 2019.

M. S. Winslett. *Reasoning about action using a possible models approach*. 1988.