Explanations as Model Reconciliation via Probabilistic Logical Reasoning

Stylianos Loukas Vasileiou[1], William Yeoh[1], Tran Cao Son[2], and Alessandro Previti[3]

[1] Washington University in St. Louis, St. Louis MO, USA
{v.stylianos,wyeoh}@wustl.edu
[2] New Mexico State University, Las Cruces NM, USA
tson@cs.nmsu.edu
[3] Ericsson Research, Stockholm, Sweeden
alessandro.previti@ericsson.com

Abstract. The model reconciliation problem is a popular paradigm within the explainable AI planning community that has been proposed as a way to provide explanations from an agent to a human user about a particular plan. Existing methods to solve this problem have been restricted to planning scenarios in which the user has a deterministic model of the problem domain (i.e., users have full certainty on their beliefs). In this paper, we propose a general logic-based explanation generation framework that extends the model reconciliation problem to users with probabilistic models of the problem domain (i.e., users have varying levels of certainty about their beliefs).

Keywords: Explanations as Model Reconciliation · Probabilistic Logical Reasoning · Degrees of Belief.

1 Introduction

The *Model Reconciliation Problem* (MRP) [3] is a popular paradigm within the realm of *Explainable AI Planning* (XAIP) that integrates mental models of users[4] in the explanation generation process of an agent. These explanations bring the model of the user closer to the agent's model by transferring a minimum number of updates from the agent's model to the user's model. While many advancements towards the model reconciliation problem realize the inception of good XAIP systems [12,2], they are usually built on a facile assumption: They assume the users have deterministic models of the problem domain, i.e., users' beliefs are of Boolean attributes. However, users might disregard explanations as unconvincing if they do not conform with or reflect to some level their personal beliefs (e.g., confirmation bias).

To this extent, we are interested in generalizing MRP to account for users with probabilistic models of the problem domain, i.e., users have varying levels of certainty about their beliefs. We build on the theoretical foundations laid by [14] and extend a general logic-based framework, where given a knowledge base KB_a (of an agent) that

[4] A mental model is just the user's version of the problem which the agent possess, and interestingly, it can be expressed as a graph, a planning model, or even a logic program.

entails a formula ϕ and a knowledge base KB_h (of a user that the agent has) that does not entail ϕ, the goal is to identify an explanation ϵ for ϕ from KB_a for KB_h such that when it is used to update KB_h, then the updated KB_h has a higher *degree of belief* in ϕ.

The notion of degree of belief in this paper has a subjective (or Bayesian) interpretation and is interchangeable with the notion of degree of probability.[5] As such, degrees of belief, also referred to as *subjective beliefs*, are degrees of certainty, or credences of subjects (i.e., agents or human users), which are used to quantify the strengths of their belief attitudes.[6] A formalization of this idea would have the degree of belief in a proposition to be confined on a scale from 0 to 1, where 0 indicates absolute certainty in the falsity of the proposition, 0.5 indicates that the proposition is just as likely to be true or false, and 1 indicates absolute certainty in its truth.

2 Preliminaries

We assume basic familiarity with standard methods in propositional logic (i.e., SAT and (weighted) model counting) (see [1]), and classical planning (i.e., STRIPS and planning as SAT) (see [8]). In what follows, we assume a consistent knowledge base KB comprising propositional formulae.

Definition 1 (Skeptical Entailment). *A formula ϕ is* skeptically entailed *by KB, denoted by $KB \models^s \phi$, if $MOD(KB) \neq \emptyset$ and $\phi \in m$ for every $m \in MOD(KB)$, where $MOD(KB)$ denotes the satisfiable models of KB. Also, $KB \models^s \phi$ iff $KB \wedge \neg\phi$ is unsatisfiable.*

Definition 2 (Credulous Entailment). *A formula ϕ is* credulously entailed *by KB, denoted by $KB \models^c \phi$, if $MOD(KB) \neq \emptyset$ and $\phi \in m$ for some $m \in MOD(KB)$.*

Definition 3 (x-Support). *Given a KB s.t. $KB \models^x \phi$, an x-support for ϕ is a subset $\epsilon \subseteq KB$ such that $\epsilon \models^x \phi$ and, $\forall \epsilon' \subset \epsilon$, we have $\epsilon' \not\models^x \phi$, where $x \in \{s, c\}$ refers to either skeptical (s) or credulous (c) entailment.*

Model Reconciliation: A *Model Reconciliation Problem* (MRP), as introduced by Chakraborti et al. [3], is defined by the tuple $\Psi = \langle \phi, \pi \rangle$, where $\phi = \langle M^a, M^a_h \rangle$ is a tuple of the agent's model $M^a = \langle D^a, I^a, G^a \rangle$ and the agent's approximation of the user's model $M^a_h = \langle D^a_h, I^a_h, G^a_h \rangle$, and π is the cost-optimal plan in M^a. A solution to an MRP is a shortest explanation ϵ s.t. when it is added to the user's model, the plan π is cost-optimal in both the agent's model and the updated user's model.

In [14] we have reformulated MRP into a logic-based account:

Definition 4 (L-MRP). *Given two knowledge bases KB_a and KB_h of the agent providing an explanation and the human receiving the explanation, respectively, such that $KB_a \models^s \phi$ and $KB_h \not\models^s \phi$, the goal of L-MRP is to find an explanation $\epsilon = \langle \epsilon^+, \epsilon^- \rangle$, where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, s.t. $\widehat{KB}^{\epsilon}_h \models^s \phi$, where $\widehat{KB}^{\epsilon}_h = (KB_h \cup \epsilon^+) \setminus \epsilon^-$*

[5] To quote De Morgan from two centuries ago: "By degree of probability, we really mean, or ought to mean, degree of belief." [5].

[6] Naively, the subjective interpretation of probability as degree of belief locates probability in a subject's mind [7].

We refer to the set of formulae ϵ as the *update* of the knowledge base KB_h, where new formulae ϵ^+ from KB_a are added and erroneous formulae ϵ^- from KB_h are removed to ensure consistency. Note that in our prior work [13], we focused on the specific task of computing an explanation ϵ s.t. $\epsilon \setminus KB_h$ is an update of minimum size.

3 Probabilistic Logical Reasoning

This section describes the notion of *degree of belief* in a formula φ with respect to a knowledge base KB, a fundamental approach introduced by [10] that combines probabilistic and logical reasoning into a unified framework, namely *probabilistic logical reasoning*. In essence, the *degree of belief* in a formula φ with respect to a deterministic knowledge base KB[7] is defined to be the conditional probability of φ given KB. The complexity of computing degrees of belief with respect to a knowledge base is polynomially related to the complexity of computing the number of models of the knowledge base, which is know to be complete in #P [10]. Formally,

Definition 5 (Degree of Belief of φ w.r.t. a Deterministic Knowledge Base). *Given a deterministic knowledge base KB and a formula φ, the degree of belief in φ given KB is $P(\varphi \mid KB) = \frac{MC(KB \wedge \varphi)}{MC(KB)}$, where $MC(KB)$ is the model count of KB.*

This means that the degree of belief in φ given KB is the fraction of models of KB in which φ evaluates to true. However, if we were to consider more realistic, real-world scenarios, then we would need to able to also represent probabilistic knowledge about the world. Such a representation constitutes the need for a *probabilistic* knowledge base:

Definition 6 (Probabilistic Knowledge Base). *A probabilistic knowledge base is a tuple $\langle KB, w \rangle$, where KB is a deterministic knowledge base, and w a weight function that assigns weights to the literals l in KB such that $0 \leq w(l) \leq 1$.*

Fundamentally, a probabilistic knowledge base is a weighted knowledge base over its literals. Specifically, for each literal l of a probabilistic knowledge base, we use $w(l)$ and $w(\neg l)$ to denote the weights of the positive and negative literals, respectively. We assume that every literal l has either (1) an indifferent weight (i.e., $w(l) = w(\neg l) = 1$), often referred to as *evidence*, that is ignored by weighted model counting algorithms for computing probabilities conditioned upon the evidence; or (2) a normal weight (i.e., $w(l) \leq 1$), where $w(l) + w(\neg l) = 1$. If every literal has weight 1, then the probabilistic knowledge base reduces to a deterministic knowledge base. Further, we say that a probabilistic knowledge base $\langle KB, w \rangle$ is inconsistent if $WMC(KB) = 0$, where $WMC(KB)$ denotes the weighted model count of KB, i.e., $WMC(KB) = \sum_{m \models^s KB} \prod_{m \models^s l} w(l)$.

We should point out that Definition 6 is semantically similar to the definition of a weighted propositional formula in logic. A weighted formula representation has been widely used to solve probabilistic inference problems by means of WMC on a variety of problems, [11,6]. Following the same terminologies, the degree of belief in a formula

[7] A deterministic knowledge base is a knowledge base in the usual sense, i.e., one that contains *no* statistical/probabilistic facts.

φ with respect to a probabilistic knowledge base reduces to computing the weighted model counting of the knowledge base:

Definition 7 (Degree of Belief of φ w.r.t. a Probabilistic Knowledge Base). *Given a probabilistic knowledge base $\langle KB, w\rangle$ and a formula φ, the probability of φ given KB is $P(\varphi \mid KB) = \frac{WMC(KB \wedge \varphi)}{WMC(KB)}$.*

For the next section, we assume that KB is defined as a tuple $\langle KB, w\rangle$, where KB is deterministic if $w(l) = 1$ for all literals $l \in KB$ and is probabilistic otherwise.

4 Explanations via Probabilistic Reasoning

We now describe our framework, which generalizes L-MRP to account for probabilistic domains. We formulate the notion of an explanation in the following setting, where, we use the term $\models^x$ for $x \in \{s, c\}$ to refer to skeptical (s) or credulous (c) entailment:

> **Explanation Generation Problem (EGP):** Given two knowledge bases KB_a and KB_h and a formula φ, where $KB_a \models^x \varphi$ and $KB_h \not\models^x \varphi$, the goal is to identify an explanation $\epsilon = \langle \epsilon^+, \epsilon^- \rangle$, where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, s.t. when it is used to *update* KB_h to $\widehat{KB}_h^{\epsilon}$, $P(\varphi \mid \widehat{KB}_h^{\epsilon}) > P(\varphi \mid KB_h)$.

An explanation in this setting would increase the degree of belief in a formula with respect to a user's knowledge base by imposing an update in their knowledge base, where new formulae ϵ^+ from KB_a are added and erroneous formulae ϵ^- from KB_h are removed as there may be contradictory formulae in KB_h and ϵ^+. Furthermore, for probabilistic knowledge bases, the explanation may be of the form $\langle \epsilon, w^{\epsilon^+} \rangle$, where w^{ϵ^+} are the weights of the explanation added to KB_h. In this case, the literals are assigned weights w.r.t. KB_h's weight function if they exist in KB_h, or w.r.t. KB_a's weight function otherwise.[8] More formally, we define the knowledge base update as follows:

Definition 8 (Knowledge Base Update). *Given a knowledge base $\langle KB_h, w^{KB_h} \rangle$ and an explanation $\langle \epsilon, w^{\epsilon^+} \rangle$, where $\epsilon = \langle \epsilon^+, \epsilon^- \rangle \subseteq KB_a \cup KB_h$, the updated knowledge base is $\langle \widehat{KB}_h^{\epsilon}, \widehat{w}^{\epsilon^+} \rangle$, where $\widehat{KB}_h^{\epsilon} = (KB_h \cup \epsilon^+) \setminus \epsilon^-$ and $\epsilon^- \subseteq KB_h \setminus \epsilon^+$ is a set of formulae that must be removed from KB_h s.t. the updated $\widehat{KB}_h^{\epsilon}$ is consistent, and for each literal $l \in \epsilon^+$, $\widehat{w}^{\epsilon^+}(l) = w^{KB_h}(l)$ if $l \in KB_h$, or $\widehat{w}^{\epsilon^+}(l) = w^{\epsilon^+}(l)$ otherwise.*

Note that we make the assumption that if a literal is not contained in a knowledge base, then it is also not contained in the language of the knowledge base. Hence, when updating a knowledge base with a new literal, we also implicitly extend its language. In what follows, and unless stated otherwise, when mentioning the update of KB_h with an explanation ϵ, we will refer to the formulae of the ϵ added to KB_h (i.e., ϵ^+).

We now define the notion of an *explanation* with respect to two knowledge bases.

[8] Note that when updating a knowledge base with new literals, we preserve the property of its weight function. However, in future work we plan to investigate this from the lens of MRP.

Definition 9 (Explanation). *Given knowledge bases KB_a and KB_h and a formula φ, assume that $KB_a \models^x \varphi$ and $KB_h \not\models^x \varphi$. Then, $\epsilon \subseteq KB_a \cup KB_h$ is an* explanation *for φ from KB_a for KB_h if $\epsilon \models^x \varphi$ and $P(\varphi \mid \widehat{KB}_h^{\epsilon}) > P(\varphi \mid KB_h)$.*

One can see that an interesting concept to explore given the definition above is that of a *maximally-convincing explanation*, that is, an explanation that yields the highest degree of belief in a given formula:

Definition 10 (Maximally-Convincing Explanation). *ϵ is a* maximally-convincing explanation *for φ from KB_a for KB_h if ϵ is an explanation according to Definition 9 and $\nexists \epsilon' \subseteq KB_a \cup KB_h$ s.t. $P(\varphi \mid \widehat{KB}_h^{\epsilon'}) > P(\varphi \mid \widehat{KB}_h^{\epsilon})$.*

Unfortunately, even computing the degree of belief in a formula w.r.t. a knowledge base is $\#P$-complete [10,4]. Additionally, maximally-convincing explanations might require a higher cognitive effort from the users to understand them, as it is more likely that such explanations would also have high cardinality. As such, maximally-convincing explanations may impose practical limitations on the agent (computing the explanation) as well as the user (parsing the explanation). Nonetheless, this can be remedied as follows: We can simply seek to find an explanation ϵ that increases the degree of belief in a given formula up to at least a user-defined bound and is of *minimal cardinality* among those set of explanations. We refer to this as a *bounded-convincing explanation*:

Definition 11 (Bounded-Convincing Explanation). *ϵ is a* bounded-convincing explanation *for φ from KB_a for KB_h if ϵ is an explanation according to Definition 9 and (*i*) $P(\varphi \mid \widehat{KB}_h^{\epsilon}) \geq b_p$, where b_p is a user-defined lower bound; and (*ii*) $\nexists \epsilon' \subset \epsilon$ s.t. ϵ' is an explanation.*

Therefore, a bounded-convincing explanation is an explanation that is *convincing*, i.e., it increases the degree of belief up to at least a desirable bound, and *concise*, i.e., it has the minimal cardinality among the set of convincing explanations.

5 Conclusions

The main purpose of this work is to provide an account for generating more personalized explanations for users. Such an account may pave the way for trustworthy autonomous agents, as users may be more inclined to accept explanations that account for their beliefs (e.g., confirmation bias). Nevertheless, the biggest challenge here would be the ability to learn the user's beliefs. Even though there has been some interest on that end [9], further research is required. We posit, though, that this may be achieved by exploiting the potential advantage of an interconnected explanation generation process. For instance, a *multi-shot explanation* approach would allow users to interact, in a collaborative manner, with the agent's explanation process, and thus allow them to provide useful information about their knowledge, beliefs, or preferences.

To conclude, we proposed a general logic-based framework for the *model reconciliation problem* (MRP), which generalizes it to account for users with varying levels of confidence in their beliefs. Due to its logical nature, our framework has the advantage of being able to deal with problems coming from different settings, so long as the problems can be represented with a logical formalism.

References

1. Biere, A., Heule, M., van Maaren, H.: Handbook of Satisfiability (2009)
2. Chakraborti, T., Sreedharan, S., Kambhampati, S.: Balancing explicability and explanations in human-aware planning. In: IJCAI. pp. 1335–1343 (2019)
3. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In: IJCAI. pp. 156–163 (2017)
4. Chavira, M., Darwiche, A.: On probabilistic inference by weighted model counting. Artificial Intelligence **172**(6-7), 772–799 (2008)
5. De Morgan, A.: Formal logic: or, the calculus of inference, necessary and probable. Taylor and Walton (1847)
6. Fierens, D., den Broeck, G.V., I. Thon, B.G., Raedt, L.D.: Inference in probabilistic logic programs using weighted cnf's. In: UAI. pp. 211–220 (2011)
7. Huber, F.: A Logical Introduction to Probability and Induction. Oxford University Press (2018)
8. Kautz, H., Selman, B.: Planning as satisfiability. In: ECAI. pp. 359–363 (1992)
9. Pynadath, D.V., Wang, N., Rovira, E., Barnes, M.J.: Clustering behavior to recognize subjective beliefs in human-agent teams. In: AAMAS. pp. 1495–1503 (2018)
10. Roth, D.: On the hardness of approximate reasoning. Artificial Intelligence **82**(1-2), 273–302 (1996)
11. Sang, T., Beame, P., Kautz, H.A.: Performing bayesian inference by weighted model counting. In: AAAI. pp. 475–481 (2005)
12. Sreedharan, S., Chakraborti, T., Kambhampati, S.: Handling model uncertainty and multiplicity in explanations via model reconciliation. In: ICAPS. pp. 518–526 (2018)
13. Vasileiou, S.L., Previti, A., Yeoh, W.: On exploiting hitting sets for model reconciliation. In: AAAI (2021)
14. Vasileiou, S.L., Yeoh, W., Son, T.C.: A preliminary logic-based approach for explanation generation. In: ICAPS XAIP Workshop (2019)