ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230543

PLEASE: Generating Personalized Explanations in Human-Aware Planning

Stylianos Loukas Vasileiou[a;*] **and William Yeoh**[a]

[a]Washington University in St. Louis

Abstract. *Model Reconciliation Problems* (MRPs) and their variant, *Logic-based MRPs* (L-MRPs), have emerged as popular methods for explainable planning problems. Both MRP and L-MRP approaches assume that the explaining agent has access to an assumed model of the human user receiving the explanation, and it reconciles its own model with the human model to find the differences such that when they are provided as explanations to the human, they will understand them. However, in practical applications, the agent is likely to be fairly uncertain on the actual model of the human and wrong assumptions can lead to incoherent or unintelligible explanations. In this paper, we propose a less stringent requirement: The agent has access to a task-specific *vocabulary* known by the human and, *if* available, a human model capturing confidently-known information. Our goal is to find a *personalized explanation*, which is an explanation that is at an appropriate abstraction level with respect to the human's vocabulary and model. Using a logic-based method called *knowledge forgetting* for generating abstractions, we propose a simple framework compatible with L-MRP approaches, and evaluate its efficacy through computational and human user experiments.

1 Introduction

Human-aware planning (HAP) is a rapidly growing area of research on helping human users interface with AI agents in complex (sequential) decision-making tasks [13]. A typical HAP scenario involves an AI agent explaining an inexplicable decision to a human user due to differences in their mental models of the task.[1] This approach is referred to as the *model reconciliation problem* (MRP) [7], and its predominant goal is to explain, in terms of model differences, why an agent's decision (e.g., a proposed sequence of actions to reach a desired goal) is valid/optimal in the agent's model but not in the human's model. A common thread around most MRP approaches is the assumption that the agent has a version of the human's model available and that it is at the same granularity level as the agent's model [6, 7, 22, 26]. Albeit a necessary assumption for establishing the foundations of MRP, it can, arguably, limit the overall practicality of MRP frameworks, insofar as the agent's version of the human model may diverge from the actual human's model and, importantly, it may be at a different abstraction level, thus leading to generating incoherent or unintelligible explanations.

However, one can argue that if the agent is not confident on its estimate of the human model, then the human model should only capture the information that the agent is confident in. The downside of this argument is that, in practical applications, the agent is likely to be fairly uncertain on most aspects of the human model and, as such, it is likely that the human model is almost empty in most cases. Consequently, the explanation generated will be unnecessarily long because the agent is falsely assuming that the human is (almost) completely unaware of any task-specific knowledge,

Therefore, in this paper, we propose that the agent *has access to a vocabulary of task-specific terms* of the human user and *generates explanations with respect to that vocabulary*. This assumption is a reasonable tradeoff between assuming that the human model is almost empty, which is overly pessimistic, and assuming that the human model is mostly specified, which is overly optimistic. Further, it can also be used in conjunction with human models that capture the information that the agent is confident the human user knows, allowing it to leverage the strength of existing MRP algorithms.

As an example, consider the classical LOGISTICS domain [18]. The vocabulary of the human may include the different trucks (e.g., `truck1`, `truck2`) and locations (e.g., `loc1`, `loc2`), and their partial model includes the action dynamics of the `move` operator for trucks. One advantage of this approach is that the vocabulary implicitly encodes the human's knowledge or expertise level of the given task. For instance, the more (or less) terms included in the vocabulary, the higher (or lower) the human's level of expertise, to the extent that a human expert probably knows more task-specific terms than a novice one, all else being equal. Continuing with the example above, the human user is knowledgeable about trucks, but is unaware of the existence of airplanes. The agent could then exploit the vocabulary and construct explanations tailored to the human's level.

To that end, we focus on *logic-based MRPs* (L-MRPs), a variant of MRPs where the underlying optimization and explanation generation problems can be encoded in a logical language [21, 25, 26]. We propose a framework, where given an agent knowledge base KB_a encoding a task, an explanandum φ entailed by KB_a, a (possibly partial or empty) human knowledge base KB_h, and a human vocabulary $\mathcal{V}_h$ consisting of task-specific terms, the goal is to find an explanation that is at an appropriate abstraction level with respect to KB_h and $\mathcal{V}_h$. To do this, we employ a fundamental logic-based operation, namely *knowledge forgetting* [2, 24], and describe how it can be used for generating abstractions. We then formally define the notion of *personalized explanations* and present an algorithm that can be combined with any off-the-shelf L-MRP approaches for computing them. Finally, we empirically demonstrate the efficacy of our framework on a set of representative benchmarks as well on a human user study.

* Corresponding Author. Email: v.stylianos@wustl.edu.
[1] In a typical planning task, the agent's and the human's (mental) models encode their own understanding of the problem's dynamics.

2 Logical Preliminaries

Throughout the paper, we assume a propositional language L consisting of a finite set of propositional letters Γ. The simplest formulae in L are *literals*, which are letters or their negations. More complex formulae can be recursively build up from letters and the classical logical connectives. A *knowledge base* KB is a set of formulae. The set of letters used in KB's formulae is called the *vocabulary* of KB, denoted by $\mathcal{V}_{KB}$. An *interpretation* is a function $\mathcal{I} : \Gamma \to \{\top, \bot\}$, and if there exists an interpretation that satisfies a KB, then KB is *satisfiable*, otherwise KB is *unsatisfiable*, denoted by $KB \models \bot$. A KB entails a formula φ, denoted by $KB \models \varphi$, iff $KB \cup \{\neg\varphi\} \models \bot$. Unless stated otherwise, we assume that a KB is satisfiable and expressed in *conjunctive normal form* (CNF), that is, a conjunction of clauses, each of which is a disjunction of literals.

Definition 1 (Explanation)**.** *Given $KB \models \varphi$, $\epsilon \subseteq KB$ is an* explanation *for φ from KB iff $\epsilon \models \varphi$ and $\forall \epsilon' \subset \epsilon$, $\epsilon' \not\models \varphi$.*

Example 1. *Let $KB = \{a, b, \neg a \vee c\}$, where $KB \models c$. Then, $\epsilon = \{a, \neg a \vee c\}$ is an explanation for c from KB.*

In this paper, we build upon the logic-based variant of model reconciliation (L-MRP) [26]. As an L-MRP explanation must take into account both the knowledge base KB_a of the agent providing an explanation as well as the knowledge base KB_h of the human receiving the explanation, an *L-MRP explanation* is defined slightly differently compared to Definition 1:

Definition 2 (L-MRP Explanation)**.** *Given knowledge bases $KB_a \models \varphi$ and $KB_h \not\models \varphi$, $\epsilon = \langle \epsilon^+, \epsilon^- \rangle$ is an* L-MRP explanation *for φ from KB_a to KB_h iff $\epsilon^+ \subseteq KB_a$, $\epsilon^- \subseteq KB_h$, and $(KB_h \cup \epsilon^+) \setminus \epsilon^- \models \varphi$.*

When KB_h is updated with an L-MRP explanation ϵ, new formulae ϵ^+ from KB_a are added and formulae ϵ^- from KB_h are removed to ensure consistency. Among the set of possible explanations, Vasileiou *et al.* [26] proposed a number of cost functions (e.g., subset-minimality, cardinality-minimality, etc.).

3 Related Work

The central area of this work is *human-aware planning* (HAP) [5,6]. A popular approach within HAP is called *model reconciliation problem* (MRP) [7, 21–23, 25, 26]. In this approach, the (planning) agent must have knowledge of the human's model in order to contemplate their goals and foresee how its plan will be perceived by them. When there exist differences between the models of the agent and the human such that the agent's plan diverges from the human's expectations, the agent provides a minimal set of model differences (i.e., an explanation) to the human.

In the introduction, we described a key limitation of existing MRP approaches, namely that it assumes that the human model is accurate. As such, on one hand, it can generate overly long explanations by pessimistically assuming that the human model is almost empty when it captures only information that it is confident about. On the other hand, it can generate incoherent explanations by optimistically (and wrongly) assuming that the human model is mostly specified when it also captures information that it is not confident about.

Nevertheless, our work in this paper is closely connected to a logic-based variant of model reconciliation (L-MRP) [26], where the underlying optimization and explanation generation problems can be encoded in a logical language. A limitation of existing L-MRP approaches, but not necessarily non-logic-based MRP approaches, is that they assume that $\epsilon^+ \subseteq KB_a$ must be subsets of *exact clauses* from KB_a (See Definition 2). Therefore, the human user may have to learn and understand a very complicated ϵ^+ with many new terms and concepts when a simpler version with fewer new terms could have sufficed. For example, in classical planning, only one action is allowed to be executed at each time step. A logic-based encoding of this restriction is through the following rule:

$$\bigwedge_{a \in A} \bigwedge_{a' \in A | a \neq a'} (\neg a_t \vee \neg a'_t) \tag{1}$$

where A is the set of actions in the problem and a_t represents action a in timestep t. Note that A is the set of *all actions* in the problem. As such, should the human user be unaware of this rule, ϵ^+ would include it and, thus, they need to learn about all possible actions.

Continuing with the LOGISTICS example in the introduction, imagine that the agent is trying to explain that it is not possible to execute, at the same time, both actions `move(truck1, loc1, loc2)` and `move(truck1, loc1, loc3)`, which correspond to moving the truck to both locations `loc2` and `loc3` concurrently. An explanation generated through existing L-MRP methods will include Rule 1 with the set of *all actions* when a much simpler and shorter version where A is composed of *exactly the two* `move` *actions* above only would have sufficed. Our goal in this paper is to enrich the L-MRP formulation by proposing a new algorithmic framework that can find such simpler explanations through knowledge forgetting.

In the context of classical planning problems, Sreedharan *et al.* [23] considered a related issue. They assumed the user's model is part of an abstraction lattice held by the agent, with each node representing an abstracted planning task model produced by projecting out a set of state fluents. The agent estimates the appropriate level based on user interactions and provides consistent explanations. This is achieved through a foil set (a set of plans) provided by the user, which the agent uses to find a minimal set of models consistent with the foil. This method can be seen as a special case of MRP, with the user model belonging to a set of possible models representing various abstractions of the agent's model. In contrast, our approach follows the standard MRP assumption with a single estimate of the user's model and a human-specified vocabulary that may include terms not in the user's model. This allows us to capture scenarios where users have relevant vocabulary terms without knowledge of their relationship to the problem. This can often be the case when the problem includes terms, such as `move` in LOGISTICS, that are in everyday conversation. Additionally, our approach has the merit of generality, as it can be applied to problems beyond planning, as long as they can be encoded in a logical formalism for which satisfiability of sets is feasible.

Finally, with recent progress, large language models (LLMs) [1] may be able to tackle explainability and reconciliation challenges. As few-shot learners, LLMs excel at producing well-formed sentences [3,16]. Nevertheless, their primary shortcomings in establishing a robust basis for logical reasoning, mainly due to their dependence on statistical features for inference, has been exploited [8,20]. Conversely, our framework's symbolic nature offers important theoretical guarantees, such as logically consistent and accurate explanations. The ability to perform consistent logical reasoning is vital for building trust between human users and AI systems.

4 Knowledge Forgetting

Knowledge forgetting, henceforth forgetting, has an ordering function in the human mind – it can be seen as a process of omitting infor-

mation or knowledge from one's memory in such a way that it is no longer present or reproducible. From a cognitive point of view, forgetting is a gradual process in which information that is less used is moved to the "background," from which it either dissipates or recovered through remembering to the foreground [10]. This basic mechanism helps people deal with information overload by suppressing irrelevant information, which allows them to focus on the relevant aspects of a given task, thus improving their cognitive capabilities. For example, when people are trying to focus on a specific task, they tend to "forget" irrelevant aspects around it, or when trying to find a solution under restricted conditions, they have to intentionally "forget" ways of solving the task in more granular environments [12]. This point to the fact that intentional forgetting is a fundamental cognitive process involving many aspects of knowledge and reasoning.

Interestingly, the operation of forgetting aligns with a pragmatic framework in cognitive linguistics called *relevance theory* [28]. Relevance theory suggests that the relevance of a statement transmitted to an individual should minimize their cognitive effort (i.e., effort in processing the statement) and maximize their positive cognitive effect (i.e., the statement leads to a true conclusion). In other words, the more positive the cognitive effects and the less the cognitive effort, the greater the relevance of the statement to the individual. The connection we ought to draw here is that forgetting can be seen as an operation for achieving the objectives suggested by relevance theory, so far as forgetting irrelevant information from a statement can decrease the individual's effort, and by focusing only on what is relevant, yield a positive effect. In the sequel, we look at forgetting from the lens of logic and show how it can be used for that purpose.

4.1 Logic-based View of Knowledge Forgetting

Analogous to the cognitive operation of forgetting, which aims at suppressing information from an agent's memory, the logic-based operation of forgetting aims at removing information from an agent's knowledge base. Forgetting has received many logical definitions and interpretations, starting in the mid 1800s with Boole's variable elimination method [2]. For a historical overview of forgetting in logic and AI, we refer the reader to the work by Van Ditmarsch *et al.* [24].

Generally, forgetting is defined through an operation that decreases the language of an agent, insofar as the vocabulary of the agent's language is reduced. Intuitively, forgetting information from an agent's knowledge base that encodes a specific domain affects the agent's ability to express or represent information about that domain, rather than losing information about the domain per se.

Delgrande [9] presents a resolution-based mechanism for computing forgetting, where given a knowledge base KB defined over vocabulary $\mathcal{V}_{KB}$, the operation of forgetting $\lambda \subseteq \mathcal{V}_{KB}$ from KB is the logical consequences of KB expressible over $\mathcal{V}_{KB} \setminus \lambda$.

Given a knowledge base KB and a letter $\lambda \in \mathcal{V}_{KB}$ in its vocabulary, let $KB^{\downarrow\lambda}$ and $KB^{\uparrow\lambda}$ denote the sets of formulae of KB that do not mention λ and do mention λ, respectively:

$$KB^{\downarrow\lambda} = \{\varphi \in KB \mid \lambda \notin \mathcal{V}_\varphi\} \tag{2}$$

$$KB^{\uparrow\lambda} = \{\varphi \in KB \mid \lambda \in \mathcal{V}_\varphi\} \tag{3}$$

Additionally, let $Res(KB^{\uparrow\lambda}, \lambda)$ denote the set of formulae obtained from $KB^{\uparrow\lambda}$ by carrying out all possible resolutions with respect to letter λ:

$$Res(KB^{\uparrow\lambda}, \lambda) = \{\varphi \mid \exists \varphi_1, \varphi_2 \in KB^{\uparrow\lambda} \text{ s.t. } \lambda \in \varphi_1, \neg\lambda \in \varphi_2, \varphi = (\varphi_1 \setminus \{\lambda\}) \cup (\varphi_2 \setminus \{\neg\lambda\})\} \tag{4}$$

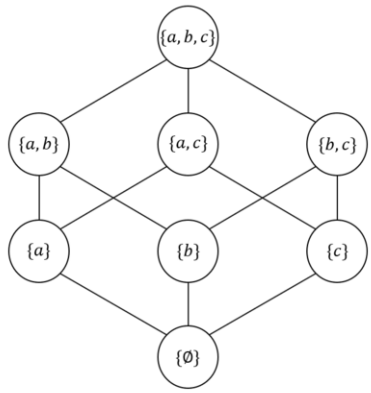

Figure 1: Abstraction lattice for $KB = \{a, b, \neg a \vee c, \neg b \vee \neg c \vee d\}$. At the root is level-0 of the lattice, i.e., the initial $\mathcal{F}(KB, \{\emptyset\}) = KB$. The child nodes of the root form level-1 of the lattice and represent (from left to right): $\mathcal{F}(KB, \{a\}) = \{b, c, \neg b \vee \neg c \vee d\}$, $\mathcal{F}(KB, \{b\}) = \{a, \neg a \vee c, \neg c \vee d\}$, and $\mathcal{F}(KB, \{c\}) = \{a, b, \neg a \vee \neg b \vee d\}$. Similarly, the subsequent nodes form level-2, and so on.

Combining those definitions, we get the definition of forgetting:

Definition 3 (Forgetting)**.** *Given a knowledge base KB and a letter $\lambda \in \mathcal{V}_{KB}$ in its vocabulary,* forgetting *λ from KB is defined as $\mathcal{F}(KB, \lambda) = KB^{\downarrow\lambda} \cup Res(KB^{\uparrow\lambda}, \lambda)$.*[2]

Definition 3 can be interpreted as follows: Perform all possible resolutions with respect to the letter to be forgotten, and add these resolvents to those formulae in KB that do not mention that letter. While the resulting KB is weaker than before as it loses its expressivity with respect to what is forgotten, one key advantage is that it still entails the same set of formulae that are irrelevant to what was forgotten:

Property 1. *If $KB \models \varphi$, then $\forall \lambda \subseteq \mathcal{V}_{KB} \setminus \mathcal{V}_\varphi$, $\mathcal{F}(KB, \lambda) \models \varphi$.*

Example 2. *Let $KB = \{a, b, \neg a \vee c, \neg b \vee \neg c \vee d\}$ with $\mathcal{V}_{KB} = \{a, b, c, d\}$. Notice that $KB \models d$. For $\lambda = \{a\}$, we get $KB^{\downarrow a} = \{b, \neg b \vee \neg c \vee d\}$ and $KB^{\uparrow a} = \{a, \neg a \vee c\}$, and $Res(KB^{\uparrow a}, a) = \{c\}$. Then, $\mathcal{F}(KB, \{a\}) = \{b, c, \neg b \vee \neg c \vee d\}$, where $\mathcal{F}(KB, \{a\}) \models d$.*

Abstractions via Forgetting: As seen from the example above, the forgetting operation can be viewed as a method for simplifying formulae by "forgetting" a set of letters. In essence, if we define an abstraction of a knowledge base as simplifying it, then forgetting is a succinct operation for computing various abstraction levels:

Definition 4 (Abstraction)**.** *Given a knowledge base KB and a set of letters $\lambda \subseteq \mathcal{V}_{KB}$ in its vocabulary, a level-$|\lambda|$* abstraction *of KB is $\mathcal{F}(KB, \lambda)$.*

We can now create an *abstraction lattice* defining the abstraction levels that can be achieved on a knowledge base given a set of letters. Figure 1 shows a level-3 abstraction lattice based on Example 2. As we will see in the next section, generating personalized explanations boils down to finding the appropriate abstraction level with respect to a set of letters (i.e., the human-specified vocabulary).

[2] Note that computing forgetting for a set of letters can be done iteratively (i.e., $\mathcal{F}(KB, \lambda_1 \cup \lambda_2) = \mathcal{F}(\mathcal{F}(KB, \lambda_1), \lambda_2)$).

5 Personalized Explanation Generation

Our framework builds upon the *logic-based model reconciliation problem* (L-MRP) [26], where we make the following assumptions:

- The agent has a knowledge base KB_a encoding its knowledge of a task in a logical language. The agent's knowledge base KB_a is logically closed, insofar as the agent is "logically omniscient" about the task.
- The agent has a knowledge base KB_h encoding, possibly incompletely or erroneously, the human user's knowledge of the same task in the same logical language. It is possible for $KB_h = \emptyset$.
- The human user provides to the agent: (*i*) An explanandum φ, where $KB_a \models \varphi$ and $KB_h \not\models \varphi$, and (*ii*) a vocabulary $\mathcal{V}_h$. Naturally, $\mathcal{V}_{KB_h} \subseteq \mathcal{V}_h$ as all the terms in the human model must be in their vocabulary. However, note that *it is possible for the vocabulary to include terms that are not in the human model*. This is akin to knowing a particular term, but not knowing how it relates to the task or what it really means.

Thus, given the knowledge bases KB_a and KB_h, the corresponding human vocabulary $\mathcal{V}_h$, and an explanandum φ such that $KB_a \models \varphi$ and $KB_h \not\models \varphi$, the goal is to find an L-MRP explanation from KB_a to KB_h for φ that is *at an appropriate abstraction level with respect to* $\mathcal{V}_h$. As already mentioned, we will call such an explanation a *personalized explanation*.

The central question behind this setting is what is an appropriate abstraction level. Clearly, an appropriate abstraction level should not contain any irrelevant information with respect to the explanandum:

Definition 5 (Irrelevance)**.** *Given a knowledge base $KB_a \models \varphi$, a set of letters $\lambda \subseteq \mathcal{V}_{KB_a}$ from its vocabulary is* irrelevant *for KB_a with respect to φ iff $\mathcal{F}(KB_a, \lambda) \models \varphi$.*

We say that a set of letters λ is irrelevant for KB_a with respect to a formula φ if forgetting λ from KB_a does not affect the entailment of φ in the resulting KB_a. In our context, this definition is easily satisfied by assuming that λ does not contain any letters from the explanandum φ (see Property 1). We enforce this property in our proposed algorithm, which we describe later.

Further, a personalized explanation is not really "personalized" unless it uses at least some letters familiar to the human (i.e., letters from the vocabulary $\mathcal{V}_h$). Naively, one could forget all letters from $\mathcal{V}_{KB_a}$ except for those in $\mathcal{V}_h$ (and $\mathcal{V}_\varphi$). However, this may result in overly short and trivial explanations of the form "why φ? Because φ", which is the case when $KB_h = \emptyset$ and $\mathcal{V}_h \cap \mathcal{V}_{KB_a} = \emptyset$.

Therefore, to avoid forgetting too many letters and oversimplifying explanations to the point that they are trivial, we propose that the goal of forgetting as many letters as needed to get an (L-MRP) explanation is of reasonable complexity. The complexity of explanations can be measured in a variety of ways, including with all the various cost functions (e.g., subset-minimality, cardinality, etc.) previously proposed in the literature [26].

Without loss of generality, we will assume that the choice of complexity measure is the cardinality of the explanation. While we continue our description, provide theoretical properties, and propose an algorithm based on this assumption, it is fairly straightforward to see how they can be generalized to other complexity measures as well.

When the choice of complexity measure is explanation cardinality, the constraint that needs to be satisfied is:

$$|\mathcal{F}(\epsilon^+, \lambda)| + |\epsilon^-| \leq \mathcal{UB} \tag{5}$$

where $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_h \cup \mathcal{V}_\varphi)$ is the set of letters to forget and $\mathcal{UB}$ is a user-specific maximum cardinality of an explanation. Note that the λ does not include any letters in the vocabulary $\mathcal{V}_h$ because the goal is to personalize explanations by using terms known to the human user. Additionally, λ does not include any letters in $\mathcal{V}_\varphi$ because they are needed to ensure that the updated KB_h of the user entails the explanandum φ (Property 1). Finally, no letters are forgotten from $\epsilon^- \subseteq KB_h$ because they are all in the vocabulary $\mathcal{V}_h$ of the human user by definition. More formally, extending Definition 2:

Definition 6 (Personalized L-MRP Explanation)**.** *Given knowledge bases $KB_a \models \varphi$ and $KB_h \not\models \varphi$, vocabulary $\mathcal{V}_h$, and upper bound $\mathcal{UB}$, $\epsilon = \langle \epsilon^+, \epsilon^- \rangle$ is a* personalized L-MRP explanation *for φ from KB_a to KB_h with respect to $\mathcal{V}_h$ iff $\epsilon^+ \subseteq KB_a$, $\epsilon^- \subseteq KB_h$, $\lambda \in \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h)$, $|\mathcal{F}(\epsilon^+, \lambda)| + |\epsilon^-| \leq \mathcal{UB}$, and $(KB_h \cup \mathcal{F}(\epsilon^+, \lambda)) \setminus \epsilon^- \models \varphi$.*

Given Definitions 2 and 6 together with Property 1, we can then show that if $\langle \epsilon^+, \epsilon^- \rangle$ is an L-MRP explanation for φ from KB_a to KB_h, then $\langle \mathcal{F}(\epsilon^+, \lambda), \epsilon^- \rangle$ is a personalized L-MRP explanation for φ from KB_a to KB_h for any $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h)$ if its cardinality is no larger than a given upper bound $\mathcal{UB}$. More formally:

Theorem 1. *Given two knowledge bases $KB_a \models \varphi$ and $KB_h \not\models \varphi$ with a corresponding L-MRP explanation $\langle \epsilon^+, \epsilon^- \rangle$ for explanandum φ, for any set of letters $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h)$ and an upper bound $\mathcal{UB}$, $\langle \mathcal{F}(\epsilon^+, \lambda), \epsilon^- \rangle$ is a personalized L-MRP explanation for the same explanandum φ if its cardinality $|\mathcal{F}(\epsilon^+, \lambda)| + |\epsilon^-| \leq \mathcal{UB}$ is no larger than $\mathcal{UB}$.*

Proof. Assume $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h)$, which is the premise of the theorem. Then:

$$\begin{aligned}
\mathcal{F}((KB_h \cup \epsilon^+) \setminus \epsilon^-, \lambda) &= \mathcal{F}((KB_h \setminus \epsilon^-) \cup (\epsilon^+ \setminus \epsilon^-), \lambda) && (6)\\
&= \mathcal{F}((KB_h \setminus \epsilon^-) \cup \epsilon^+, \lambda) && (7)\\
&= \mathcal{F}(\epsilon^+ \cup (KB_h \setminus \epsilon^-), \lambda) && (8)\\
&= \mathcal{F}(\epsilon^+, \lambda) \cup (KB_h \setminus \epsilon^-) && (9)\\
&= (\mathcal{F}(\epsilon^+, \lambda) \cup KB_h) \setminus \epsilon^- && (10)\\
&= (KB_h \cup \mathcal{F}(\epsilon^+, \lambda)) \setminus \epsilon^- \models \varphi && (11)
\end{aligned}$$

The simplification from Equations 6 to 7 is due to the properties of L-MRP explanations that $\epsilon^- \subseteq KB_h$, $\epsilon^+ \subseteq KB_a$, and that the intersection $KB_h \cap KB_a$ will never be part of the explanation since that information is already common to both knowledge bases (Definition 2). The simplification from Equations 8 to 9 is because $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h)$ (premise of the theorem) does not contain any letters in KB_h or its subset $\epsilon^- \subseteq KB_h$. For the same reason, Equation 9 can be rewritten as Equation 10. Finally, the entailment in Equation 11 is because $\mathcal{F}((KB_h \cup \epsilon^+) \setminus \epsilon^-, \lambda)$ entails φ since $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h) \subseteq \mathcal{V}_{(KB_h \cup \epsilon^+) \setminus \epsilon^-} \setminus \mathcal{V}_\varphi$ (Property 1). Combining this entailment and the premise that $|\mathcal{F}(\epsilon^+, \lambda)| + |\epsilon^-| \leq \mathcal{UB}$, we can conclude that $\langle \mathcal{F}(\epsilon^+, \lambda), \epsilon^- \rangle$ is a personalized L-MRP explanation (Definition 6). □

5.1 Computing Personalized Explanations

Our algorithm, called *Personalized Logical Explanation Algorithm for Symbolic Environments* (PLEASE), exploits Theorem 1 to find personalized L-MRP explanations. Algorithm 1 describes its pseudocode. At a high level, PLEASE is composed of the following steps:

(1) Use any off-the-shelf L-MRP solver to find a sequence of L-MRP explanations.

Algorithm 1: Personalized Logical Explanation Algorithm for Symbolic Environments (PLEASE)

Input: Agent knowledge base KB_a, human knowledge base KB_h, explanandum φ, human vocabulary $\mathcal{V}_h$, upper bound $\mathcal{UB}$
Result: A *personalized explanation* $\langle \epsilon^+, \epsilon^- \rangle$

```
while true do
    ⟨ε+, ε−⟩ ← next-L-MRP-exp(KB_a, KB_h, φ)
    if ⟨ε+, ε−⟩ = null then
        return null
    else
        foreach λ ⊆ V_ε+ \ (V_h ∪ V_φ) do
            ε+ ← F(ε+, λ)
            if |ε+| + |ε−| ≤ UB then
                return ⟨ε+, ε−⟩
```

(2) For each L-MRP explanation $\langle \epsilon^+, \epsilon^- \rangle$, iterate through all possible subsets of letters $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h)$.

(3) For each subset of letters λ to forget, check if the cardinality of the explanation with forgotten letters $|\mathcal{F}(\epsilon^+, \lambda)| + |\epsilon^-| \leq \mathcal{UB}$ is within the upper bound $\mathcal{UB}$.

(4) If it is, then return the personalized explanation. If not, repeat with the next L-MRP explanation from the L-MRP solver.

Example 3. *Let* $KB_a = \{a, d, \neg d \vee b, \neg a \vee \neg b \vee c\}$, $KB_h = \emptyset$, $\varphi = \{c\}$, *and* $\mathcal{V}_h = \{a, d\}$, *and suppose that we are searching for a personalized explanation whose cardinality is within an upper bound* $\mathcal{UB}$ *of 3. First, notice that since* $KB_h = \emptyset$, *the only L-MRP explanation is* $\epsilon^+ = KB_a$ *and* $\epsilon^- = \emptyset$. *As* $|\epsilon^+| + |\epsilon^-| = 4 + 0 = 4$ *is larger than the upper bound, we will iterate through all possible subsets* $\lambda \subseteq \mathcal{V}_{\epsilon^+} \setminus (\mathcal{V}_\varphi \cup \mathcal{V}_h) = \{a, b, c, d\} \setminus (\{c\} \cup \{a, d\}) = \{b\}$, *which in this case is only the letter* b. PLEASE *then checks if forgetting this letter is sufficient to reduce the cardinality of the explanation to within the upper bound:* $|\mathcal{F}(\epsilon^+, \lambda)| + |\epsilon^-| = |\mathcal{F}(\epsilon^+, b)| = |\{a, d, \neg a \vee \neg d \vee c\}| = 3$. *Since it is,* PLEASE *will return the personalized explanation* $\langle \{a, d, \neg a \vee \neg d \vee c\}, \emptyset \rangle$.

It is fairly straightforward to see that the algorithm is correct and complete, under the assumption that the underlying off-the-shelf L-MRP solver is also correct and complete.

6 Empirical Evaluations

We now empirically evaluate our approach both in simulated computational experiments as well as in a human user study.

6.1 Simulated Computational Experiments

We ran the experiments on a MacBook Pro machine comprising an M1 Max processor with 32GB of memory. The time limit was set to 300s. PLEASE was implemented in Python, where we use the algorithm described by Vasileiou *et al.* [25] as the off-the-shelf solver to find L-MRPs.[3] We used our own implementation for the knowledge forgetting operation.[4]

We encoded some classical planning problems from the International Planning Competition (IPC) in the style of Kautz *et al.* [14], and used them to form the agent's knowledge base KB_a. The explanandum φ for each problem was the plan optimality query, which we constructed as described in [25]. We varied three parameters – the assumed knowledge base of the human KB_h, the vocabulary of the human $\mathcal{V}_h$, and the upper bound $\mathcal{UB}$. To construct the knowledge base KB_h, we follow the literature by modifying KB_a, specifically, by removing p fraction of actions as well as p fraction of preconditions and effects of each remaining action. To construct the vocabulary $\mathcal{V}_h$, we first extract all the letters that are used in KB_h, then supplement it with letters from KB_a if $|\mathcal{V}_h| \leq q$ fraction of $|\mathcal{V}_{KB_a}|$. Finally, we parameterize the upper bound $\mathcal{UB}$ as a fraction r of the cardinality of the shortest L-MRP explanation $|\epsilon^*|$. The default values for our three parameters are as follows: $p = 0.8$ for KB_h, $q = 0.4$ for $\mathcal{V}_h$, and $r = 0.8$ for $\mathcal{UB}$.

In our first experiment, we vary the completeness of KB_h by varying $p \in \{0.2, 0.4, 0.6, 0.8\}$. Table 1 tabulates the results, where we report the length of an optimal plan $|\pi^*|$, the cardinality of the shortest L-MRP explanations $|\epsilon^*|$ returned by the off-the-shelf L-MRP solver, the cardinality of the personalized L-MRP explanations $|\epsilon|$ returned by PLEASE, the number of letters forgotten $|\lambda|$, and the runtimes t of PLEASE. We make the following observations: Unsurprisingly, $|\epsilon^*|$ increases as p increases. The reason is that KB_h decreases as p increases. Therefore, more information needs to be provided in the explanation in order for the updated KB_h to entail the explanandum. Additionally, $|\lambda|$ increases as p increases as well because more needs to be forgotten from longer explanations in order to get their cardinality to within $\mathcal{UB}$. Finally, as $|\pi^*|$ increases, the cardinality of both explanations $|\epsilon^*|$ and $|\epsilon|$ increases. Consequently, the runtime t increases as well.

In our second experiment, we vary the size of the vocabulary $|\mathcal{V}_h|$ by varying $q \in \{0.2, 0.4, 0.6, 0.8\}$. Table 2 tabulates the results. As q (and, equivalently, the vocabulary size $|\mathcal{V}_h|$) increases, $|\lambda|$ decreases and $|\epsilon|$ increases since fewer letters need to be forgotten before the updated KB_h entails the explanandum. Additionally, the cardinality of the L-MRP explanation $|\epsilon^*|$ and runtimes t remain relatively unchanged for all values of q. This implies that the runtime of PLEASE is dominated by the time needed to find the L-MRP explanation by the off-the-shelf solver, and the time needed to personalize the explanations is relatively small.

In our third experiment, we vary the upper bound $\mathcal{UB}$ by varying $r \in \{0.6, 0.7, 0.8, 0.9\}$. Table 3 tabulates the results. As r (and, equivalently, the upper bound $\mathcal{UB}$) increases, $|\lambda|$ decreases and $|\epsilon|$ increases since fewer letters need to be forgotten to get a personalized explanation with a cardinality that is within the upper bound $\mathcal{UB}$. Similar to the second experiment, the cardinality of the L-MRP explanation $|\epsilon^*|$ and runtimes t remain relatively unchanged for all values of r, making the same implication that the runtime of PLEASE is dominated by the off-the-shelf solver.

6.2 Human User Study

We now evaluate one of the assumptions made in this paper, namely, that personalized explanations with respect to a human vocabulary increase the overall comprehension and satisfaction of human users. It is important to note that while we could not control for the knowledge of the human user (i.e., KB_h), it is reasonable to assume that their knowledge grows as their vocabulary $\mathcal{V}_h$ grows. For example, in the LOGISTICS domain example, if a user's vocabulary includes the `move` operator, even if they do not know specifically the preconditions and effects of the operator, they still have an intuitive sense of what it does. As such, they would have a larger KB_h compared to

[3] We used the implementation provided by the authors, which is publicly available at https://github.com/vstylianos/aaai21.
[4] The code repository is available at https://github.com/YODA-Lab/PLEASE.

Prob.		π^*	$p=0.2$				$p=0.4$				$p=0.6$				$p=0.8$			
			$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t
BLOCK WORLD	1	4	3	2	1	0.1s	3	2	1	0.1s	54	43	12	1.5s	54	43	12	1.0s
	2	10	3	2	1	0.4s	4	3	1	0.5s	17	14	3	2.5s	17	13	3	2.5s
	3	18	10	8	1	0.5s	20	15	3	1.5s	59	44	19	33.0s	59	46	25	37.0s
LOGIS-TICS	1	12	5	4	1	1.0s	5	4	1	1.0s	10	8	2	2.0s	12	10	4	3.0s
	2	14	4	3	1	2.0s	6	5	1	3.5s	8	6	2	4.0s	11	8	2	5.5s
	3	20	6	5	1	24.5s	6	5	1	25.0s	10	8	2	25.0s	13	10	3	26.0s
TPP	1	5	16	13	3	1.5s	16	13	3	2.0s	16	12	4	1.5s	16	13	3	0.1s
	2	18	43	34	9	1.5s	43	34	9	1.5s	43	34	9	1.5s	43	34	9	1.4s
	3	27	85	68	17	144.5s	85	68	17	144.0s	85	68	17	145.0s	85	68	17	145.0s
DEPOT	1	2	5	3	31	0.5s	6	4	1	1.0s	14	11	3	2.0s	14	11	3	1.2s
	2	6	7	5	1	1.0s	7	5	1	1.5s	14	11	3	32.5s	14	11	3	37.5s
	3	10	11	8	2	2.5s	12	7	3	3.0s	26	21	5	185.0s	26	21	5	184.0s

Table 1: Evaluation of PLEASE with different completeness of knowledge bases $|KB_h|$.

Prob.		π^*	$q=0.2$				$q=0.4$				$q=0.6$				$q=0.8$			
			$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t
BLOCK WORLD	1	4	54	43	12	1.2s	54	43	12	1.0s	54	43	12	1.1s	54	43	11	1.0s
	2	10	17	12	5	3.0s	17	13	3	2.5s	17	13	4	3.0s	17	14	3	2.5s
	3	18	59	47	12	41.0s	59	46	25	37.0s	59	44	13	42.0s	59	44	15	43.0s
LOGIS-TICS	1	12	12	8	5	4.0s	12	10	4	3.0s	12	10	2	3.0s	12	10	3	3.0s
	2	14	11	8	2	5.0s	11	8	2	5.5s	10	8	2	5.2s	10	8	2	5.0s
	3	20	13	9	3	27.5s	13	10	3	26.0s	13	9	4	25.5s	13	10	3	26.0s
TPP	1	5	16	13	3	0.1s	16	13	3	0.1s	16	13	3	0.1s	16	13	3	0.1s
	2	18	43	34	9	1.5s	43	34	9	1.4s	43	34	11	2.0s	43	34	9	1.6s
	3	27	85	68	17	143.0s	85	68	17	145.0s	75	68	17	145.0s	85	67	17	146.0s
DEPOT	1	2	14	11	3	1.2s	14	11	3	1.2s	14	11	3	1.2s	14	11	3	1.0s
	2	6	14	11	3	38.0s	14	11	3	37.5s	14	11	3	37.5s	14	11	3	38.0s
	3	10	26	21	5	186.0s	26	21	5	184.0s	26	21	5	187.0s	26	21	5	185.5s

Table 2: Evaluation of PLEASE with different sizes of vocabulary $|V_h|$.

Prob.		π^*	$r=0.6$				$r=0.7$				$r=0.8$				$r=0.9$			
			$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t	$\lvert\epsilon^*\rvert$	$\lvert\epsilon\rvert$	$\lvert\lambda\rvert$	t
BLOCK WORLD	1	4	54	32	22	1.3s	54	38	16	1.3s	54	43	12	1.0s	54	49	5	1.2s
	2	10	17	10	7	2.3s	17	12	5	2.5s	17	13	3	2.5s	17	15	2	2.5s
	3	18	59	22	33	37.0s	59	22	33	37.5s	59	46	25	37.0s	59	53	6	37.0s
LOGIS-TICS	1	12	12	7	5	3.5s	12	8	4	3.5s	12	10	4	3.0s	12	11	1	3.0s
	2	14	11	7	4	6.0s	11	8	3	5.0s	11	8	2	5.0s	11	9	1	4.5s
	3	20	13	8	5	29.0s	13	9	4	29.0s	13	10	3	26.0s	13	12	1	29.0s
TPP	1	5	16	10	6	0.04s	16	11	5	0.1s	16	13	3	0.1s	16	14	2	0.1s
	2	18	43	26	18	1.5s	43	30	13	1.4s	43	34	9	1.4s	43	30	4	1.5s
	3	27	85	51	34	133.0s	85	59	26	135.0s	85	68	17	145.0s	85	76	9	136.5s
DEPOT	1	2	14	8	6	1.2s	14	10	4	1.5s	14	11	3	1.2s	14	13	1	1.0s
	2	6	14	8	6	35.0s	14	10	4	34.0s	14	11	3	37.0s	14	13	1	35.0s
	3	10	26	16	9	179.0s	26	18	7	180.0s	26	21	5	184.0s	26	23	2	186.5s

Table 3: Evaluation of PLEASE with different upper bounds $\mathcal{UB}$.

another user who does not know about the operator, everything else being equal. With this in mind, our hypothesis is that:

> *Human users with access to a known vocabulary of task-specific terms (and their background knowledge associated to those terms) have an increased understanding and satisfaction with personalized explanations compared to human users with generic explanations.*

Study Design: We designed a between-subject user study, wherein the users were divided into three vocabulary group pairs ($\mathcal{V}_{h1}$, $\mathcal{V}_{h2}$, and $\mathcal{V}_{h3}$), each of which consists of a treatment group and a control group. The study comprised a simple imaginary scenario that involved a robot exploring an environment, and a supervisor (i.e., the users) observing its behavior from a station. For simplicity, we simulated the environment as a 5x4 grid and informed the users about the robot's capabilities, such as moving to adjacent locations among other actions. After the users understood the necessary information, they instructed the robot to move to a certain location in the grid.

Question	Vocabulary $\mathcal{V}_{h1}$			Vocabulary $\mathcal{V}_{h2}$			Vocabulary $\mathcal{V}_{h3}$		
	Treatment	Control	Sig?	Treatment	Control	Sig?	Treatment	Control	Sig?
Q1: The explanation helped me understand the robot's decision to communicate the data.	**3.90**	2.65	**Yes**	**4.45**	2.60	**Yes**	**3.60**	2.70	**Yes**
Q2: I am satisfied with the robot's explanation about how it behaved.	**3.90**	2.70	**Yes**	**4.40**	2.65	**Yes**	**3.65**	2.90	**Yes**
Q3: I feel that the explanation of how the robot behaved has sufficient detail.	**3.75**	2.60	**Yes**	**3.90**	2.90	**Yes**	**3.65**	2.80	**Yes**
Q4: I feel that the explanation of how the robot behaved is complete.	**3.75**	2.60	**Yes**	**3.95**	2.80	**Yes**	**3.15**	2.80	No
Q5: I found the robot's explanation useful for understanding its behavior.	**3.70**	2.40	**Yes**	**4.10**	2.65	**Yes**	**3.50**	2.80	No
Q6: I am confident in my understanding of the explanation.	**4.15**	2.30	**Yes**	**4.30**	2.65	**Yes**	**3.45**	2.75	No
Q7: I am confident in my ability to explain the robot's behavior (based on its explanation) to someone else.	**3.95**	2.10	**Yes**	**4.05**	2.30	**Yes**	**3.35**	2.75	No

Table 4: Average scores (max. score 5) and statistical significance (t-test, $p = 0.05$) on each question in the treatment and control groups.

To generate explanations, we told the users that on top of moving to the particular location, the robot also communicated some data to their station and, as supervisors, they requested an explanation so as to understand its behavior. The explanations were in natural language; however, some of the terms in the explanation were changed to random Greek letters. These letters then formed the three vocabulary groups (i.e., $\mathcal{V}_{h1}$, $\mathcal{V}_{h2}$, and $\mathcal{V}_{h3}$, which has one, two, and three letters with meanings described, respectively). Within each vocabulary group, the treatment group received a personalized explanation, where explanations were provided using only the vocabulary known to the group, whereas the control group received the default explanation without any personalization.[5]

The main task of the users was to evaluate the robot's explanation. To do this, we asked the users seven Likert-type questions pertaining to the understandability and satisfaction of the explanation.

Results: In total, we recruited 120 users (40 for each vocabulary group pair, 20 in the treatment and 20 in the control group) from the online crowdsourcing platform Prolific [19], with the only filter being that the users are fluent in English. Table 4 tabulates the average scores for each Likert-type question and whether the scores are statistically significant with respect to a t-test based on a p-value of 0.05. The distributions of all questions can be found in the supplement.

The results presented in Table 4 show a clear trend in favor of personalized explanations. When comparing the treatment and control groups for each vocabulary, we observe that the treatment group consistently scores higher on average across the seven Likert-type questions. This indicates that personalized explanations tailored to users' known vocabulary can lead to an increased understanding and satisfaction compared to generic explanations.

In the case of vocabulary groups $\mathcal{V}_{h1}$ and $\mathcal{V}_{h2}$, the treatment group outperforms the control group in all questions, with statistical significance observed at a $p = 0.05$ level. This suggests that personalizing explanations based on a smaller, more focused vocabulary (i.e., one or two terms) has a considerable impact on users' understanding and satisfaction. These results support our hypothesis that personalized explanations can be more effective than generic ones when users have access to a known vocabulary of task-specific terms.

However, when we examine the results for vocabulary group $\mathcal{V}_{h3}$, we notice that the treatment group only shows statistically significant improvements in Q1, Q2, and Q3. This finding may suggest that as the vocabulary size increases, the benefits of personalized explanations become less pronounced. Further investigation is needed to better understand this relationship and its implications on the design of personalized explanations.

In summary, the results of our study demonstrate the value of personalized explanations for enhancing user understanding and satisfaction, especially when a smaller, focused vocabulary is used. While the effectiveness of personalization appears to decrease with larger vocabularies, the overall trend suggests that tailoring explanations to users' known vocabulary can lead to better outcomes than providing generic explanations. Future research could explore the potential trade-offs between vocabulary size and personalization to better understand the optimal conditions for delivering effective explanations.

[5] For additional details about the study design and how the explanations were generated in accordance to our framework, see the supplement available at https://github.com/YODA-Lab/PLEASE.

7 Conclusions

In this paper, we looked at generating explanations at appropriate abstraction levels with respect to a human vocabulary via a method called knowledge forgetting. While the operation of knowledge forgetting has been extensively studied in various logical settings [11, 17, 27, 30], its applicability in the context of human-aware planning and explanation generation has not been fully explored, to the best of our knowledge. We hope this work adds to the growing research on human-aware planning, enabling effective communication and collaboration between humans and AI agents, while ensuring a coherent and personalized explanation experience.

It is important to note that in addition to explanation generation, *explanation communication* is a crucial aspect of explanatory systems that is often ignored. It has been shown that *explanations as model reconciliation*, presented mostly as text, serve an important and intuitive way of explaining plans to users [4, 29]. Additionally, Kumar *et al.* [15] showed that conveying explanations through visualizations tend to be more preferred by users than text alone. On that premise, and given the logical nature of our framework, we ought to say that we do not aim at presenting explanations to users in a logical format. The final form of our explanations can be, for instance, translated into natural language before communicated to a user. We will pursue this in future work.

At the other end of the spectrum, we view the work presented here as a necessary step towards realizing an interactive, multi-shot explanation generation scheme, where human users interact with an agent in a dialogical fashion. The personalized explanations presented here can serve as a foundation for instigating a dialogue between the user and the agent. Specifically, we conceptualize a framework where, upon receiving an initial explanation from the agent, the user would have the option to request further clarifications by pointing to specific parts of the explanation, in which case the agent will increase (or decrease) the explanation's granularity. Another option for the user would be to refute the agent's explanation (i.e., engage in an argumentative process). Through these interactions and the information exchange, the agent will be able to update their approximation of the user's knowledge base and, thus, learn more accurate representations of the user's actual knowledge. We hope to pursue this exciting direction in the future as well.

Acknowledgments

This research is partially supported by the National Science Foundation under awards 1812619 and 2232055. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring organizations, agencies, or the United States government.

References

[1] Rishi Bommasani, Drew Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., 'On the opportunities and risks of foundation models', *arXiv preprint arXiv:2108.07258*, (2021).

[2] George Boole, *An Investigation of the Laws of Thought: On which are Founded the Mathematical Theories of Logic and Probabilities*, Dover, 1854.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., 'Language models are few-shot learners', in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1877–1901, (2020).

[4] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati, 'Plan explanations as model reconciliation–an empirical study', in *Proceedings of the International Conference on Human-Robot Interaction (HRI)*, pp. 258–266, (2019).

[5] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati, 'Balancing explicability and explanations in human-aware planning', in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1335–1343, (2019).

[6] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati, 'The emerging landscape of explainable automated planning & decision making.', in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4803–4811, (2020).

[7] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati, 'Plan explanations as model reconciliation: Moving beyond explanation as soliloquy', in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 156–163, (2017).

[8] Antonia Creswell, Murray Shanahan, and Irina Higgins, 'Selection-inference: Exploiting large language models for interpretable logical reasoning', in *Proceedings of the International Conference on Learning Representations (ICLR)*, (2023).

[9] James Delgrande, 'A knowledge level account of forgetting', *Journal of Artificial Intelligence Research*, **60**, 1165–1213, (2017).

[10] Hermann Ebbinghaus, 'Memory: A contribution to experimental psychology', *Annals of Neurosciences*, **20**(4), 155, (2013).

[11] Thomas Eiter and Gabriele Kern-Isberner, 'A brief survey on forgetting from a knowledge representation and reasoning perspective', *Künstliche Intelligenz*, **33**(1), 9–33, (2019).

[12] Hollyn Johnson, 'Processes of successful intentional forgetting.', *Psychological Bulletin*, **116**(2), 274, (1994).

[13] Subbarao Kambhampati, 'Synthesizing explainable behavior for human-AI collaboration', in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1–2, (2019).

[14] Henry Kautz and Bart Selman, 'Planning as satisfiability', in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pp. 359–363, (1992).

[15] Ashwin Kumar, Stylianos Loukas Vasileiou, Melanie Bancilhon, Alvitta Ottley, and William Yeoh, 'VizXP: A visualization framework for conveying explanations to users in model reconciliation problems', in *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, pp. 701–709, (2022).

[16] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch, 'Pretrained transformers as universal computation engines', in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, p. 7628–7636, (2022).

[17] Carsten Lutz and Frank Wolter, 'Foundations for uniform interpolation and forgetting in expressive description logics', in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 989–995, (2011).

[18] Drew McDermott, 'The 1998 AI planning systems competition', *AI Magazine*, **21**(2), 35–35, (2000).

[19] Stefan Palan and Christian Schitter, 'Prolific. ac—a subject pool for online experiments', *Journal of Behavioral and Experimental Finance*, **17**, 22–27, (2018).

[20] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al., 'Scaling language models: Methods, analysis & insights from training gopher', *arXiv preprint arXiv:2112.11446*, (2021).

[21] Tran Cao Son, Van Nguyen, Stylianos Loukas Vasileiou, and William Yeoh, 'Model reconciliation in logic programs', in *Proceedings of the European Conference on Logics in Artificial Intelligence (JELIA)*, pp. 393–406, (2021).

[22] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati, 'Handling model uncertainty and multiplicity in explanations via model reconciliation', in *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, pp. 518–526, (2018).

[23] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati, 'Using state abstractions to compute personalized contrastive explanations for AI agent behavior', *Artificial Intelligence*, **301**, 103570, (2021).

[24] Hans Van Ditmarsch, Andreas Herzig, Jérôme Lang, and Pierre Marquis, 'Introspective forgetting', *Synthese*, **169**(2), 405–423, (2009).

[25] Stylianos Loukas Vasileiou, Alessandro Previti, and William Yeoh, 'On exploiting hitting sets for model reconciliation', in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6514–6521, (2021).

[26] Stylianos Loukas Vasileiou, William Yeoh, Tran Cao Son, Ashwin Kumar, Michael Cashmore, and Daniele Magazzeni, 'A logic-based explanation generation framework for classical and hybrid planning problems', *Journal of Artificial Intelligence Research*, **73**, 1473–1534, (2022).

[27] Yisong Wang, Yan Zhang, Yi Zhou, and Mingyi Zhang, 'Knowledge forgetting in answer set programming', *Journal of Artificial Intelligence Research*, **50**, 31–70, (2014).

[28] Deirdre Wilson and Dan Sperber, *Relevance Theory*, Blackwell, 2002.

[29] Zahra Zahedi, Alberto Olmo, Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati, 'Towards understanding user preferences for explanation types in model reconciliation', in *Proceedings of the International Conference on Human-Robot Interaction (HRI)*, pp. 648–649, (2019).

[30] Yan Zhang and Yi Zhou, 'Knowledge forgetting: Properties and applications', *Artificial Intelligence*, **173**(16-17), 1525–1537, (2009).