DR-HAI: Argumentation-based Dialectical Reconciliation in Human-AI Interactions

Abstract

In this paper, we introduce DR-HAI (Dialectical Reconciliation in Human-AI Interactions), a novel game-theoretic framework designed to extend model reconciliation approaches for enhanced human-AI interaction. By adopting a multi-shot reconciliation paradigm and not assuming a-priori knowledge of the human user's model, DR-HAI enables interactive dialogues to address knowledge discrepancies between explainee and explainer agents. We provide formal operational semantics for DR-HAI using logic-based argumentation and offer theoretical guarantees regarding the framework's termination and success. Furthermore, we conduct a human-user study that compares DR-HAI to single-shot reconciliation approaches, demonstrating the efficacy of our framework in improving users' understanding of AI decisions in tasks characterized by substantial knowledge asymmetry. Our findings suggest that DR-HAI offers a promising direction for fostering effective human-AI interactions.

1 Introduction

The significance of creating AI agents that can establish trust and accountability by interacting with human users is constantly increasing. This has been the central idea behind *human-aware AI problems* (HAIP), which aim to design agents that are transparent and explainable by considering the human user's knowledge, preferences, and values (Sreedharan, Kulkarni, and Kambhampati 2022). A crucial aspect of HAIP is human-AI interaction, where the AI agent is expected to communicate effectively with humans to enhance their understanding and address any concerns. Most research efforts within HAIP have been placed on (sequential) decision-making tasks (e.g., planning, scheduling, or reinforcement learning) involving an agent and a human user, and the goal is to make the agent's decisions explainable and transparent to the human user when those decisions appear inexplicable to them.

While there have been a plethora of approaches to solve HAIP from different perspectives (Chakraborti, Sreedharan, and Kambhampati 2020), a process called *model reconciliation* has recently attracted increasing attention (Chakraborti et al. 2017; Sreedharan, Chakraborti, and Kambhampati 2018; Sreedharan et al. 2019; Son et al. 2021; Vasileiou, Previti, and Yeoh 2021; Vasileiou et al. 2022). In model reconciliation, it is assumed that the agent and the human user each have their own (mental) models of the task,[1] and the need for explainability arises due to some knowledge asymmetry between these two models that makes the agent's decision inexplicable with respect to the human user's model. A solution in model reconciliation is then an explanation (technically, a minimal set of model updates) from the agent to the human user such that the agent's decision becomes explicable to the human user.

Despite the wide success of model reconciliation approaches, we identify two limitations. First, it is assumed that the agent possesses the human user's model a-priori in order to anticipate their goals and predict how its decision will be perceived by them. This can lead to incorrect assumptions about the human user's knowledge and preferences, and consequently to unsatisfactory explanations. Further, most model reconciliation approaches are formulated around a single-shot reconciliation paradigm, that is, they focus on generating a single, albeit comprehensive, explanation that is presented all at once. While this type of reconciliation can be useful when the human user needs to quickly understand a decision or when the underlying task is relatively simple, it will fail to work for more complex decisions and tasks that require a more detailed understanding from the human.

Our proposal in this paper is to extend the model reconciliation process by proposing a new framework that is formulated around a multi-shot reconciliation paradigm with no assumptions about an a-priori human user model. Particularly, we make use of the notion of *dialectical reconciliation*,[2] and propose a framework aimed at enhancing human-AI interaction by helping an explainee *understand* an agent's decision and/or behavior (see Section 5 for our notion of understanding in this context). Our framework, called *dialectical reconciliation in human-AI interactions* (DR-HAI), is formalized in a game-theoretic manner, wherein an explainee agent and an explainer agent make moves in a game governed by rules that define the appropriateness of their utterances. We describe the *operational semantics* of DR-

[1] A model in model reconciliation represents the associated agent's understanding of the task, e.g., their knowledge about the task and each other's capabilities.

[2] Dialectic refers to a discourse involving more than one individuals who hold differing views on a subject, with the goal of determining the truth through logical argumentation.

HAI, i.e., the underlying mechanism for selecting appropriate moves, with the use of logic-based argumentation (Besnard and Hunter 2001), thus making it possible for the participants to express their conflicting views through formal arguments, and as such, reconcile their differences. Additionally, we provide some theoretical guarantees pertaining to the *termination* and *success* of DR-HAI.

In short, our thesis in this paper is premised as follows:

> *For tasks involving a higher degree of complexity, intricacy, or difficulty (non-trivial tasks) and substantial knowledge discrepancies between the explainee and explainer, DR-HAI is better suited than single-shot MRP because it improves the human user's understanding of the agent's decisions and behavior through an interactive dialogue, leading to more effective human-AI interactions.*

In the next section, we evaluate the practical feasibility of our thesis and report findings from a human-user study, focusing on the role of DR-HAI in fostering human-AI interaction and understanding in the context of tasks with knowledge discrepancies between the explainee and explainer.

2 Empirical Findings: Single-Shot Reconciliation vs Dialectical Reconciliation

In this section, we present a human-user study designed to investigate the advantages of dialectical reconciliation (DR-HAI) compared to single-shot model reconciliation (MRP) in explaining complex scenarios to users.

Our study is centered around a scenario where a human user tries to understand and troubleshoot an AI home assistant robot, called "Roomie," which appears to be disconnected. The user is provided with several prompts, such as checking the associated mobile app, confirming Roomie's connection to the charging base, verifying Roomie's connection to the internet via a wired connector, and noting a flashing light next to the LAN port. Unknown to the user, several issues complicate the situation, including an outdated mobile app, an expired license for the wired connection, and a low battery indicated by the flashing light.

We selected this scenario due to its relatability to individuals who have encountered product issues, as well as due to the complexity involved in comprehending causal relationships and contributing factors. From a technical standpoint, this narrative allowed us to approximate the human model accurately, facilitating the use of a single-shot model reconciliation-based method as a baseline.

2.1 Human-user Study Design

Participants were introduced to the problem through a narrative dialogue that explained the scenario's premise and known information. After posing the initial query (e.g., "Why are you disconnected?"), participants were divided into two groups:

> **MRP**: The first group received a single-shot MRP explanation, where the human model was assumed to include the information provided during the scenario's introduction.

	MRP	**DR-HAI**	**Filtered DR-HAI**
# of Participants	49	48	37
Comprehension score (max. score 4)	0.30	2.60	**2.94**
Satisfaction score (max. score 5)	2.94	3.56	**3.73**

Table 1: Results for the user study.

> **DR-HAI**: The second group interacted with DR-HAI explanations, choosing from four unique questions (counterarguments to Roomie's responses) in a game-like format. They could continue asking questions or decide to end the interaction.

Upon completing their interaction with Roomie, participants were asked four multiple-choice questions to evaluate their understanding of the issues, generating a comprehension score. They also responded to four Likert-scale questions (1: strongly disagree, 5: strongly agree) to gauge their satisfaction with the interaction and explanations, resulting in a satisfaction score.

Our study's hypothesis is:

> **H1:** The DR-HAI group will achieve higher comprehension and satisfaction scores than the single-shot MRP explanation group.

2.2 Study Results and Discussion

We conducted the user study using the Prolific crowdsourcing platform, recruiting 100 participants, of whom 97 completed the study. All participants were proficient in English and had at least an undergraduate education. They were given a base payment of $2.50 and the possibility of earning a $2.00 bonus for correctly answering the comprehension questions.

In the DR-HAI group, some participants chose to end the interaction after asking only one question. We filtered out these participants, creating a third subgroup called "filtered-DR-HAI" with a total of 37 participants. The study results are shown in Table 1, displaying the average scores for comprehension and satisfaction questions.

As anticipated, the MRP participants scored lower on comprehension questions, possibly due to their inability to ask follow-up questions and only receiving information based on Roomie's assumed model of them. In contrast, the DR-HAI participants outperformed the MRP group, with the filtered group (participants that asked more than one question) achieving the highest scores. This suggests that engaging in dialectical interactions and having the ability to ask contrasting questions enhances the users' overall understanding of the issues.

A similar conclusion can be drawn from the satisfaction scores, which indicate that the DR-HAI group was more satisfied due to their dialectical interaction with Roomie. Therefore, the study results support hypothesis **H1**, demonstrating the advantages of dialectical reconciliation in complex explanation scenarios.

3 Background: Logic-based Argumentation

We now review part of an existing proposal for logic-based argumentation (Besnard and Hunter 2001). To simplify the presentation of our framework in this paper, we consider a propositional language $\mathcal{L}$ with the classical entailment relation denoted by $\models$. We use to denote tautology, and $\bot$ to denote falsity, as usual. Unless stated otherwise, we assume that a knowledge base KB (a finite set of formulae) is consistent.

We adopt an intuitive notion of a logical *argument*. Essentially, a logical argument is a set of formulae that can be used to (classically) prove some claim, where each claim is represented by a formula.

Definition 1. *Let KB be a knowledge base and φ a formula. An argument for φ from KB is defined as $A = \langle \Phi, \varphi \rangle$ s.t.: (i) $\Phi \subseteq KB$; (ii) $\Phi \models \varphi$; (iii) $\Phi \not\models \bot$; and (iv) $\nexists \Phi' \subset \Phi$ s.t. $\Phi' \models \varphi$.*[3]

We call φ the **claim** of the argument, denoted by $Claim(A)$, and Φ the **premise** of the argument, denoted by $Premise(A)$. The set of all arguments for a claim φ from KB is denoted by $\mathcal{A}(KB, \varphi)$. We may sometimes refer to $Premise(A)$ as the **support** for φ.

Example 1. *Assume $KB = \{a, b, a \wedge b \rightarrow c, g, g \rightarrow a\}$. Then, an argument for c from KB is $A_1 = \langle \{a, b, a \wedge b \rightarrow c\}, c \rangle$. Another argument for c from KB is $A_2 = \langle \{b, g, g \rightarrow a \wedge b \rightarrow c\}, c \rangle$*

Recall that a dialectical approach presupposes that there is conflicting knowledge between the participating agents. As such, we now conceptualise the notion of a *counterargument*, that is, an argument that disagrees with another argument by highlighting points of contention. We consider a common counterargument type, namely **undercuts**, and define it with respect to two knowledge bases.

Definition 2. *Let KB_i and KB_j be two knowledge bases, and let $A_i = \langle \Phi, \varphi \rangle$ and $A_j = \langle \Psi, \psi \rangle$ be two arguments for φ from KB_i and for ψ from KB_j, respectively. We say that $A_i = \langle \Phi, \varphi \rangle$ is an undercut for $A_j = \langle \Psi, \psi \rangle$ iff $\varphi = \neg(\psi_1 \wedge \ldots \wedge \psi_n)$ s.t. $\{\psi_1, \ldots, \psi_n\} \subseteq \Psi$.*

Example 2. *Assume $KB_i = \{a, b, a \wedge b \rightarrow c\}$ and $KB_j = \{l, d, l \wedge d \rightarrow \neg b, e, e \rightarrow \neg c\}$, and let $A_i = \langle \{a, b, a \wedge b \rightarrow c\}, c \rangle$ be an argument for c from KB_i. Then, $A_j = \langle \{l, d, l \wedge d \rightarrow \neg b, \}, \neg b \rangle$ is an* undercut *for A_i from KB_j.*

We will refer to undercuts simply as *conflicts*, and use $\mathcal{C}(KB, A)$ to denote the set of all conflicts for an argument A from KB.

The following proposition states that if two disjointly consistent knowledge bases are jointly inconsistent, then there exists at least one pair of conflicts between them.

Proposition 1. *If $KB_i \cup KB_j \models \bot$, then there exists at least one conflict from KB_i (or KB_j) for an argument from KB_j (or KB_i).*

[3]The minimality constraint ensures that arguments remain relevant by avoiding unnecessary premises and identifying precise reasons for the inference of a claim. It also prevents any negative effects caused by superfluous premises.

As we will see in the next section, logic-based argumentation forms the underlying machinery of our proposed framework.

4 The DR-HAI Framework

We now present the *dialectical reconciliation in human-AI interactions* framework (DR-HAI). We assume that there are two agents and two roles, an *explainee* indicated by index e and an *explainer* indicated by index r, engaging into a dialogue. The scope of the dialogue is to assist the explainee in *understanding* the decisions made by the explainer, and ultimately, the explainer's behavior.[4] We will use ϕ to denote an explainer's decision, and φ to denote the set of all decisions the explainee wants to understand. Stated differently, the primary high-level goal of DR-HAI is for the explainee to understand the explainer's reasoning from which φ follows.

We first highlight three important assumptions underlying DR-HAI:

- *Agent knowledge bases:* The explainee is associated with knowledge base KB_e that represents their understanding of the explainer's knowledge of the task. The explainer is associated with knowledge base KB_r that represents its own knowledge of the task. Neither agent has a-priori access to the other agent's knowledge base.
- *Explainee queries:* The explainee has a set of queries $\phi \in \varphi$ for the explainer,[5] where $KB_e \not\models \phi$ or $KB_e \models \neg\phi$, and $KB_r \models \phi$. The explainee initiates a dialogue with an initial query ϕ.
- *Commitment stores:* Each agent has access to a *commitment store*, defined as a tuple $CS_x = \langle CS_x^1, \ldots, CS_x^t \rangle$ ($x \in \{e, r\}$), that stores each agent's utterances throughout the dialogue. Both agents have access to each other's commitment stores during the dialogue.[6]

Now, the primary goal of DR-HAI can be formulated as follows:

> *Given an explainee with KB_e and an explainer agent KB_r and a set of queries φ s.t. $\forall \phi \in \varphi$, $KB_e \not\models \phi$ (or $KB_e \models \neg\phi$) and $KB_r \models \phi$, the goal of DR-HAI is to* enable *$KB_e \models \phi$ for all ϕ through a dialogue.*

The most crucial part in this formulation is how to effectively enable $KB_e \models \phi$ during the course of a dialogue between the explainee and the explainer. In a higher level sense, what we seek to do is find a way that helps the explainee transition from a state of non-understanding a decision ϕ (e.g., $KB_e \not\models \phi$ or $KB_e \models \neg\phi$) to a state of understanding the decision (e.g., $KB_e \models \phi$). As we have already emphasized, our thesis in this paper is that a natural way of

[4]While there may exist different modes of understanding (e.g., visual, and textual understanding), the notion of understanding we adopt is tied to the notion of *explanation*. See Section 5 for more.

[5]Note that the explainee can dynamically generate subsequent queries in the course of the dialogue.

[6]A commitment store is akin to a "chat log" and is used to store all the information that has been exchanged between the agents in the dialogue.

achieving this transition is via an argumentation-based dialogue that facilitates *dialectical reconciliation.*

4.1 The Dialectical Reconciliation Dialogue Type

We now formulate the dialectical reconciliation dialogue type by adopting the tradition of Hamblin's dialectical games (Hamblin 1970, 1971). Under this model, a dialogue is formulated in a game-theoretic manner, where utterances are viewed as moves in a game governed by rules that define their applicability. We can conceptualize the moves as consisting of a set of *locutions* that define the types of utterances agents are permitted to make. Taking into account the goal of DR-HAI, we define the following set of locutions:

$$L = \{query, support, refute, agree\text{-}to\text{-}disagree\} \quad (1)$$

The $query$ locution allows the explainee to ask the explainer for an argument, the $support$ locution allows the explainer to assert an argument that supports the explainee's query, the $refute$ locution allows both agents to assert a counterargument, and the $agree\text{-}to\text{-}disagree$ locution allows both agents to "agree" with each other's utterances when no further queries or counterarguments can be made.

It is important to emphasize that we chose to have an $agree\text{-}to\text{-}disagree$ locution instead of a simply $agree$ (or $accept$) locution in order to indicate that the agents acknowledge each other's utterances. Ultimately, the goal of DR-HAI is not to convince the explainee about the topic φ, but rather to help them understand φ. Therefore, an $agree\text{-}to\text{-}disagree$ locution fosters a more flexible approach to discussion, wherein the agents do not have to agree with each other; they only have to acknowledge each other's utterances.

Additionally, note that we make two restrictions: (1) the $query$ locution is only available to the explainee; and (2) the $support$ locution is only available to the explainer. These are reasonable restrictions given the goal of DR-HAI, nonetheless, future work will look into relaxing them.

Now, the set of locutions are typically instantiated with specific formulae in order to make up the range of possible **dialogue moves**. A dialogue move is defined as:

$$m_k = \langle x, l, \Phi \rangle, \quad (2)$$

where k is an index indicating the timestep of the dialogue, $x \in \{e, r\}$ indicates the agent making the move, $l \in L$ is a locution, and $\Phi \in \mathcal{L}$ is a formula that instantiates the locution (e.g., the content of the move). We denote with $\mathcal{M}$ the set of all dialogue moves.

We now define the notion of a DR-HAI dialogue. A DR-HAI dialogue requires that the first move must always be a $query$ locution from the explainee, and the agents take it in turns to receive and make moves. Formally,

Definition 3. *A DR-HAI dialogue D^t is a sequence of moves $[m_1, \ldots, m_t]$, where $t \in \mathbb{N}$, involving an explainee agent e and an explainer agent r, and the following conditions hold:*

1. *$m_1 = \langle e, query, \Phi \rangle$ is the opening move of the dialogue made by the explainee.*
2. *Each agent can make and receive only one move per timestep t.*

We will call the explainee queries φ made in the dialogue as the *topic* of the dialogue. The *outcome* of the dialogue is the collective sequence of moves made during the dialogue, denoted by $O(D^t) = [m_1, \ldots, m_t]$. We use $\mathcal{D}$ to denote the set of all DR-HAI dialogues.

A DR-HAI dialogue D^t *terminates* at t if and only if the explainee cannot generate subsequent queries or counterarguments, that is, when the explainee utters the $agree\text{-}to\text{-}disagree$ locution. More formally,

Definition 4. *Let D^t be DR-HAI dialogue. We say that D^t terminates at t iff $m_t = \langle ee, agree\text{-}to\text{-}disagree, \emptyset \rangle$ and $\nexists s$ s.t. $s < t$ and D^t terminates at s.*

During the dialogue, the agents have to essentially decide which dialogue moves to make. An agent may have a specific objective in mind when making its decision – this could include adhering to rationality principles, ending the dialogue quickly, or prolonging the dialogue as much as possible. The mechanism for deciding a move that takes this overall aim into account is called an *agent strategy*. We can think of the agent strategy for an agent x as being a function $S_x : \mathcal{D} \times KB_x \rightarrow \mathcal{M}$ that takes a dialogue and the agent's knowledge base KB_x, and returns a dialogue move. While agent strategies can take several forms (e.g., preference-based, probabilistic), for simplicity, we consider two ordered strategies for the agents:

Definition 5. *Let D^t be a DR-HAI dialogue. The DR-HAI agent strategy function for the explainee i is:*

$$S_i(D^t, KB_e) = \begin{cases} 1 : \langle e, refute, \Phi \rangle \\ 2 : \langle e, query, \Phi \rangle \\ 3 : \langle e, agree\text{-}to\text{-}disagree, \emptyset \rangle \end{cases}$$

and for the explainer agent j is:

$$S_j(D^t, KB_r) = \begin{cases} 1 : \langle r, refute, \Phi \rangle \\ 2 : \langle r, support, \Phi \rangle \\ 3 : \langle r, agree\text{-}to\text{-}disagree, \emptyset \rangle \end{cases}$$

Finally, if the agents follow their agent strategies during the DR-HAI dialogue, and the dialogue does not continue after it has terminated, then we say that the DR-HAI dialogue is *well-formed*.

Definition 6. *Let D^t be a DR-HAI dialogue. We say that D^t is well-formed iff for all $1 < s < t$, D^s does not terminate at s and for each move m_s from agent x, $S_x(D^s, KB_x) = m_s$.*

For the remainder of the paper, we assume that all DR-HAI dialogues are well-formed. In what follows, we describe the operational semantics of a DR-HAI dialogue.

4.2 The Semantics of DR-HAI Dialogues

At large, not all dialogue moves are valid during a dialogue, that is, the agents cannot freely combine locutions and formulae in order to generate a move. The set of valid moves is usually specified by a *dialogue protocol*.

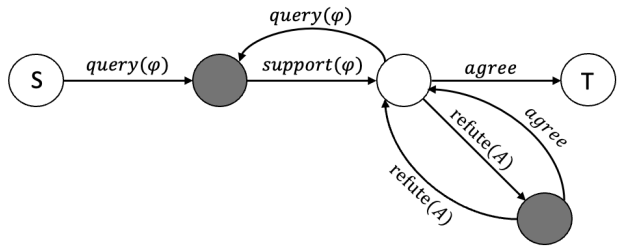

Figure 1: The DR-HAI dialogue protocol drawn as a finite-state machine. The start state is indicated by **S** and termination state is indicated by **T**. States shown without fill are states in which the explainee is expected to make a move in the dialogue; states filled in grey are states in which the explainer is expected to make a move.

Fundamentally, a dialogue protocol describes the *operational semantics* of the dialogue by specifying a set of preconditions and effects for each available locution (Plotkin 1981). What this means is that the locutions have properties of *actions*, insofar as they can change the state of the world, i.e., the state of the dialogue. Figure 1 illustrates this concept in a finite-state machine representation.

Table 2 describes how to generate valid dialogue moves m_t ($t > 1$) with respect to a DR-HAI dialogue protocol from the perspective of the explainee. A subsequent $query$ locution is instantiated with formula Φ if and only if Φ has been uttered by the explainer and has not been queried in a previous timestep (preconditions (1)-(2)) and KB_e either does not entail Φ or it entails the negation of Φ (precondition (3). The $support$ locution is instantiated with formula Φ if and only if Φ has been queried by the explainee in the previous timestep (precondition (1)) and there exists an argument for Φ from KB_r that has not been uttered before (precondition (2)). The $refute$ locution is instantiated with Φ if and only if Φ has been asserted by the other agent at any timestep in the dialogue (precondition (1)) and there exists a counterargument A for Φ from KB_e (precondition (2)).[7] Additionally, the $agree\text{-}to\text{-}disagree$ locution is uttered if and only if neither $refute$ or $query$ can be uttered.[8] Finally, the commitment stores of the agents are updated after each move.

Illustrative Example We now illustrate a DR-HAI dialogue. Assume the following knowledge bases for the explainee and the explainer:

$$KB_e = \{e, e \rightarrow \neg c, i, i \rightarrow \neg f\}$$
$$KB_r = \{a, b, a \wedge b \rightarrow c, h, h \rightarrow \neg e, f, f \rightarrow h\}$$

Additionally, assume that the explainee initiates a dialogue with an initial query $\phi = c$. The dialogue will be sequenced as follows:

$m_1 = \langle e, query, c\rangle \;\; || \;\; CS_e^1 = query(c)$

$m_2 = \langle r, support, c\rangle \;\; || \;\; CS_r^2 = \langle\{a, b, a \wedge b \rightarrow c\}, c\rangle$

$m_3 = \langle e, refute, c\rangle \;\; || \;\; CS_e^3 = \langle\{e, e \rightarrow \neg c\}, \neg c\rangle$

$m_4 = \langle r, refute, e\rangle \;\; || \;\; CS_r^4 = \langle\{h, h \rightarrow \neg e\}, \neg e\rangle$

$m_5 = \langle e, query, h\rangle \;\; || \;\; CS_e^5 = query(h)$

$m_6 = \langle r, support, h\rangle \;\; || \;\; CS_r^6 = \langle\{f, f \rightarrow h\}, h\rangle$

$m_7 = \langle e, refute, f\rangle \;\; || \;\; CS_e^7 = \langle\{i, i \rightarrow \neg f\}, \neg f\rangle$

$m_8 = \langle r, agree\text{-}to\text{-}disagree, \emptyset\rangle \;\; || \;\; CS_r^8 = agree\text{-}to\text{-}disagree$

$m_9 = \langle e, agree\text{-}to\text{-}disagree, \emptyset\rangle \;\; || \;\; CS_e^9 = agree\text{-}to\text{-}disagree$

The explainee initiates the dialogue by querying the explainer about c (m_1), and the explainer responds with an argument that supports it (m_2). The explainee then responds with a counterargument against the support (m_3), and similarly, the explainer responds back with a counterargument to the explainee's argument (m_4). The explainee then makes a subsequent query about h (m_5), for which the explainer responds with a support (m_6), and which the explainee refutes (m_7). Finally, both agents utter $agree\text{-}to\text{-}disagree$ (m_8-m_9), in which case the dialogue terminates.

4.3 Properties of DR-HAI Dialogues

We now highlight two important properties for evaluating the quality of a DR-HAI dialogue, namely, *termination* and *success*.

Termination requires that the dialogue does not run forever, and ideally, that the dialogue is gridlock-free, i.e., an agent always has a move to make at any state of the dialogue.

Theorem 1. *A DR-HAI dialogue D^t always terminates at t.*

PROOF (SKETCH). Firstly, the operational semantics (see Table 2) describe the constraints and the conditions under which each dialogue move can be executed. Secondly, the agents' knowledge bases are finite, implying that there are only a finite number of different moves that can be generated, and the agents cannot repeat these moves. Therefore, a DR-HAI dialogue D^t is guaranteed to terminate at some timestep t.

□

We now want to examine whether the terminated dialogue has been successful. At large, we can say that a dialogue is successful if the *goal* of the dialogue is attained. Recall that the primary goal of DR-HAI is for the explainee agent to understand the topic φ of the dialogue, which we formalized as $KB_e \models \phi$ for each $\phi \in \varphi$. For brevity, we will use $KB_e \models \varphi$ to denote entailment for all $\phi \in \varphi$. This can be trivially achieved by performing a *knowledge update* on KB_e with the explainer's arguments uttered in the dialogue. To do that, we adopt a simple definition for updating a knowledge base from Vasileiou et al. (Vasileiou et al. 2022):

Definition 7. *Let KB be a knowledge base and A an argument. The update on KB with A is $\widehat{KB}^A = (KB \cup Premise(A)) \setminus \gamma$, where $\gamma \subseteq KB \setminus Premise(A)$.*

[7] Notice that an agent can construct a counterargument using the other agent's commitment store.

[8] For the explainer agent, $agree\text{-}to\text{-}disagree$ is made if and only if $refute$ cannot be uttered.

Locution	Preconditions	Effects
$query(\Phi)$	(1) $\Phi \in CS_r^T$ **and** (2) $query(\Phi) \notin CS_e^T$ **and** (3) $KB_e \not\models \Phi$ or $KB_e \models \neg\Phi$	(1) $CS_e^t \leftarrow query(\Phi)$
$support(\Phi)$	(1) $query(\Phi) \in CS_e^{t-1}$ **and** (2) $\exists A \in \mathcal{A}(KB_r, \Phi)$ s.t. $A \notin CS_r^T$	(1) $CS_r^t \leftarrow A$
$refute(\Phi)$	(1) $\Phi \in CS_r^T$ **and** (2) $\exists A \in \mathcal{C}(KB_e \cup CS_r^T, \Phi)$ s.t. $A \notin CS_e^T$	(1) $CS_e^t \leftarrow A$
$agree\text{-}to\text{-}disagree$	(1) $refute(\Phi)$ preconditions do not hold **or** (2) $query(\Phi)$ preconditions do not hold	(1) $CS_e^t \leftarrow agree\text{-}to\text{-}disagree$

Table 2: The DR-HAI dialogue protocol from the perspective of explainee agent e. Recall from Definition 3 that the first move m_1 of the dialogue is always a $query$ from the explainee, and that the subsequent $query$ and $support$ locutions are only available to the explainee and the explainer, respectively. When we use CS_x^T in a condition, we ought to mean that the condition holds $\forall T \in \{1, \ldots, t-1\}$.

Definition 7 adds the premises of an argument A into the knowledge base KB, and retracts a set of formulae γ from KB if and only if KB is inconsistent with the premises added.

Note that the knowledge base update can also be performed by the explainee dynamically during the dialogue, i.e., when the explainee cannot refute any of the arguments uttered by the explainer. For simplicity, however, we assume that the update is performed at the end of the dialogue.

In our context, not all of the explainer agent's arguments would be needed to update KB_e in order for the updated KB_e to entail φ. The update could be done in an sequential fashion starting from the latest argument uttered by the explainer, and continue until the condition $KB_e \models \varphi$ is satisfied. Note that if a retraction is needed to ensure consistency, all of the previously added arguments are preserved, i.e., we only retract formulae that were in the original KB_e. This ensures that $KB_e \models \varphi$ will be satisfied.

Therefore, if the goal of the DR-HAI dialogue is met, the we say that the DR-HAI dialogue is *successful*.

Definition 8. *Let D^t be a DR-HAI dialogue on topic φ terminated at t. We say that D^t is successful iff $\widehat{KB_e}^A \models \varphi$ for some $A \subseteq CS_r$, where CS_r is the commitment store of the explainer agent.*

Finally, if we combine Definition 8 with the restrictions and assumptions underlying the overall DR-HAI framework, then we can see that a DR-HAI dialogue can be always successful.

Theorem 2. *A DR-HAI dialogue D^t on topic φ terminated at t is always successful.*

PROOF (SKETCH). Recall that the topic of the dialogue φ must be entailed by the explainer, e.g., $KB_r \models \varphi$. This means that an argument for φ from KB_r always exists (Definition 1). As in the worst case we may update the explainee's knowledge base KB_e with all arguments in the explainer's commitment store CS_r (Definition 7), it must be the case that the explainee's knowledge base will eventually entail φ, e.g., $KB_e \models \varphi$. □

5 On Explanation and Understanding

Naturally, an explanation fulfills its purpose when the explainee *understands* the information pertaining to the subject matter. The nature of an explanation, therefore, requires knowledge and understanding – it involves conveying information from an explainer in a way that is relevant and understandable to the intended explainee (Craik 1967). In other words, an explanation is about transferring knowledge from the explainer to the explainee. This transfer of knowledge is arguably most effective within the context of a dialogue, during which the explainer incrementally attains the level of specificity needed to guide the explainee toward the desired level of understanding.

Nonetheless, measuring the explainee's level of understanding can be a challenging task, as understanding is an abstract concept that may involve various factors, such as cognitive abilities. From a psychological perspective, however, understanding entails possessing a functional (mental) model of the phenomenon being explained that encompasses its causes, consequences, and related aspects. Consequently, when someone provides an explanation, what is conveyed is essentially a "blueprint" for constructing a working model (Johnson-Laird 1983). This perspective of understanding, along with substantial evidence suggesting that humans learn and understand more effectively when engaging in argumentation (Mercier and Sperber 2011), form the foundation of our proposed framework.

Since our framework is aimed at enhancing the ex-

plainee's understanding of the explainer's decisions, it would be beneficial to quantify and approximate the explainee's level of understanding. One way to do this could be by considering the similarity between the knowledge base of the explainee KB_i and the explainer KB_j.[9] Specifically, *we posit that the explainee's understanding of the explainer's decisions and behavior is likely to improve as the similarities between KB_i and KB_j increase*:

Conjecture 1. *Let $\Sigma \in [0,1]$ represent a measure of the similarity between KB_e and KB_r, and let $U \in [0,1]$ be the level of understanding the explainee has regarding the explainer. As the similarity between KB_e and KB_r increases ($\Sigma \to 1$), the level of understanding U increases ($U \to 1$).*

The conjecture posits that the explainee's knowledge base is not static, allowing it to incorporate new information. Furthermore, we assume that the explainee is a rational agent who seeks to understand the explainer's viewpoints and is willing to refine their knowledge base accordingly by integrating the explainer's information. In essence, the explainee's knowledge base continuously adapts to more closely approximate the explainer's knowledge base as additional information is shared.

Now, the similarity Σ between the two knowledge bases can be defined as a *syntactic* as well as a *semantic* similarity metric. Syntactic similarity quantifies the similarity between two knowledge bases based on their structural similarity such as the similarity of their formulae. In contrast, a semantic similarity metric gauges the similarity between knowledge bases by examining the *meaning* of their respective formulae, i.e., the logical entailments that result from the the knowledge bases.

To define a combined syntactic and similarity metric, we can employ a weighted Sørensen-Dice similarity index as follows:

$$\Sigma = \alpha \cdot \frac{2 \cdot |KB_e \cap KB_r|}{|KB_e| + |KB_r|} + (1-\alpha) \cdot \frac{2 \cdot |E_e \cap E_r|}{|E_e| + |E_r|}, \quad (1)$$

where $\alpha \in [0,1]$ is parameter indicating the weight of each metric, and E_e and E_r are the logical entailments from KB_e and KB_r, respectively. Then, the explainee's level of understanding can be approximated as the similarity between the two knowledge bases, i.e., $U \approx \Sigma$, where 0 represents no understanding, and 1 represents complete understanding of the explainer's knowledge by the explainee.

Example 3. *Assume the knowledge bases from Section 4.2. The level of understanding before the dialogue is $U \approx 0$. As we described in Section 4.3, upon termination of the dialogue the explainee will sequentially update KB_e with the explainer's arguments until the topic φ of the dialogue is entailed by KB_e, i.e., $KB_e \models h$ and $KB_e \models c$, where h and c are the two queries uttered in the dialogue. We now show how the explainee's level of understanding evolves with each knowledge update, where we used $\alpha = 0.5$ for calculating the similarity metric:*

1. *Update KB_e with premise $\{f, f \to h\}$ from CS_r^6:*
   $KB_e = \{e, e \to \neg c, i, f, f \to h\}$
   $U \approx 0.34$
2. *Update KB_e with premise $\{h, h \to \neg e\}$ from CS_r^4:*
   $KB_e = \{e \to \neg c, i, f, f \to h, h, h \to \neg e\}$
   $U \approx 0.60$
3. *Update KB_e with premise $\{a, b, a \wedge b \to c\}$ from CS_r^2:*
   $KB_e = \{e \to \neg c, i, f, f \to h, h, h \to \neg e, a, b, a \wedge b \to c\}$
   $U \approx 0.89$

6 Computational Evaluation

In this section, we present a computational evaluation of DR-HAI, utilizing the following metrics to assess its performance:

- **Dialogue Length L:** The total number of dialogue moves exchanged between the explainer and explainee agents.
- **Dialogue Time T:** The duration of the dialogue, accounting for the computational efforts required to generate arguments and counterarguments.
- **Number of Updates N:** The total count of updates to the explainee's knowledge base after the dialogue, reflecting the volume of new information absorbed by the explainee.
- **Percentage Increase in Level of Explainee Understanding I:** The percentage increase in the explainee's level of understanding, comparing their initial (pre-dialogue) and final (post-dialogue) levels.

To create explainee (KB_e) and explainer (KB_r) knowledge bases, we generated random inconsistent propositional KBs of varying sizes, resulting in 4 variants for each size: 10^2, 10^3, 10^4, totaling 12 unique KB pairs. We constructed KB_r by removing a minimal correction set (MCS) from the inconsistent KB.[10] For KB_e, we controlled the fraction of conflicts between the explainer and explainee with the ratio $|KB_e|/|KB_r| = c$. Specifically, starting with an empty KB_e, we added formulae from MCS and, if needed, negations of random formulae from KB_r to meet the desired ratio. This process generated distinct KBs with conflict levels determined by c. Lastly, to have KBs of approximately the same size and with some similarity between them, we added a $1-c$ fraction of formulae from KB_r to KB_e, as long as KB_e remained satisfiable.

For generating arguments and counterarguments, we used a method from (Besnard et al. 2010). The dialogue topic comprised a single query φ, created by finding a formula entailed by KB_r but not by KB_e. We identified such a formula by examining the logical entailments of both knowledge bases. This process ensured the query addressed the knowledge discrepancy between the explainer and explainee, allowing for a dialogue that resolved conflicts and improved the explainee's understanding.

[9] Recall from Section 4 that the explainee's knowledge base is the explainee's approximation of the explainer's knowledge.

[10] Given an inconsistent KB, a minimal correction set (MCS) represents the smallest set of formulae that makes the KB consistent when removed (Marques-Silva et al. 2013).

$\|KB\|$	Var.	$c = 0.1$				$c = 0.2$				$c = 0.3$				$c = 0.4$			
		T	L	N	I	T	L	N	I	T	L	N	I	T	L	N	I
10^2	1	0.01s	7	1	4%	0.03s	19	9	58%	0.02s	15	7	51%	0.05s	17	8	60%
	2	0.01s	5	1	18%	0.02s	15	6	45%	0.03s	11	2	33%	0.01s	7	2	32%
	3	0.05s	13	1	7%	0.01s	5	2	21%	0.03s	13	4	23%	0.03s	13	5	25%
	4	0.03s	17	5	22%	0.05s	25	3	54%	0.02s	11	2	44%	0.02s	11	2	74%
10^3	1	0.8s	11	4	2%	0.8s	11	4	7%	1.0s	25	11	17%	0.8s	9	3	6%
	2	0.5s	23	10	11%	0.2s	5	2	1%	0.1s	5	1	6%	0.2s	5	2	2%
	3	1.0s	11	5	6%	1.0s	5	2	5%	1.5s	11	5	4%	1.0s	7	3	6%
	4	0.3s	7	2	5%	3.5s	93	3	10%	0.5s	17	8	7%	1.0s	39	17	23%
10^4	1	4.0s	19	8	66%	142.0s	399	199	120%	2.0s	11	4	164%	61.0s	91	45	208%
	2	250.0s	179	11	64%	15.5s	45	22	112%	52.0s	171	1	170%	98.0s	153	76	205%
	3	400.0s	187	1	73%	300.0s	499	100	125%	171.0s	313	156	175%	496.0s	223	11	217%
	4	297.0s	143	52	103%	93.0s	184	63	193%	346.0s	313	64	194%	281.0s	239	52	301%

Table 3: Evaluation of DR-HAI on a variant of knowledge base sizes $|KB|$ and fractions of conflicts c.

6.1 Results

We implemented a prototype of DR-HAI in Python using the PySAT toolkit (Ignatiev, Morgado, and Marques-Silva 2018), and ran the experiments with a time limit of $500s$ on a MacBook Pro machine comprising an M1 Max processor with 32GB of memory.[11]

Table 3 presents the evaluation results of DR-HAI on various knowledge base sizes ($|KB|$) and fractions of conflicts (c), allowing us to observe how they influence the dialogue time (T), dialogue length (L), number of updates (N), and the improvement in the explainee's understanding (I).

The results in Table 3 reveal several trends and insights into the performance of DR-HAI:

- As the size of the knowledge base increases, the dialogue time (T) generally increases. This is expected, as larger knowledge bases require more computational resources to generate arguments and counterarguments. However, the increase is not always linear, as some instances of larger knowledge bases can be resolved more quickly than others, depending on the specific structure of the inconsistencies and the complexity of the arguments.
- The dialogue length (L) and number of updates (N) also tend to increase with larger knowledge bases and higher fractions of conflicts (c). This indicates that more dialogue moves and updates are needed to resolve inconsistencies when the knowledge bases are larger and when a higher proportion of conflicting formulae are present.
- The percentage increase in the level of explainee understanding (I) shows considerable variation across different scenarios. In general, larger knowledge bases and higher fractions of conflicts tend to result in a higher improvement in the explainee's understanding. However, there are exceptions to this trend, suggesting that other factors, such as the structure of the knowledge bases and the complexity of the inconsistencies, can also impact the improvement in understanding.

In summary, the performance of DR-HAI is influenced by various factors, including the size of the knowledge base, the fraction of conflicts, and the complexity of the inconsistencies. Future work could focus on optimizing the implementation and exploring additional factors that may impact the performance of DR-HAI.

[11]The code will be publicly made available.

7 Discussion

Our proposed framework DR-HAI is related to dialogue-based approaches. According to the influential work by Walton and Krabbe (Walton and Krabbe 1995), dialogues can be categorized based on the knowledge of the participants, the objectives they wish to achieve through the dialogue, and the rules that are intended to govern the dialogue. Contextual to each type, each dialogue revolves around a topic, typically a proposition, that is the subject matter of discussion. Indeed, with DR-HAI we are introducing a new dialogue type, namely *dialectical reconciliation*. Mostly relevant to dialectical reconciliation, are the following dialogue types: persuasion (Gordon 1994; Prakken 2006), where an agent attempts to *persuade* another agent to accept a proposition that they do not initially hold; information-seeking (Parsons, Wooldridge, and Amgoud 2003; Fan and Toni 2012), where an agent *obtains information* from another agent who is believed to possess it; and inquiry (Hitchcock and Hitchcock 2017; Black and Hunter 2009), where two agents collaborate to find a *joint proof* to a query that neither could individually. Although many dialogue systems have been proposed for the aforementioned dialogue types, to the best of our knowledge there are no existing dialogue frameworks that consider the use of dialectical reconciliation that is aimed at enhancing an explainee's *understanding*. Table 4 shows a logical description of the initial states and goal states of dialectical reconciliation and the three related dialogue types.

Our primary focus lies in addressing human-aware AI problems (Sreedharan, Kulkarni, and Kambhampati 2022), particularly model reconciliation problems (MRP) (Chakraborti et al. 2017; Sreedharan, Srivastava, and Kambhampati 2021; Son et al. 2021; Vasileiou, Previti, and Yeoh 2021; Kumar et al. 2022). Within MRP, our work closely follows the logic-based version of model reconciliation proposed by Vasileiou et al.(Vasileiou, Previti, and Yeoh 2021; Vasileiou et al. 2022), where agent and human user models are represented as logical knowledge bases. Our proposed framework addresses two key limitations of current MRP approaches: the assumption that the explainer agent knows the human user model and that interactions are single-shot.

Our proposed framework extends the model reconciliation process by incorporating a new dialogue type—dialectical reconciliation—enabling continuous interactions between the agent and the human user. This approach could foster a more comprehensive understanding for the explainee and could also promote a more collaborative and dynamic explanation process.

Limitations and Future Directions: Despite the promising aspects of our framework, it is essential to acknowledge its limitations and potential areas for improvement.

Our framework follows a fixed structure in presenting arguments and does not consider the effectiveness of personalizing the interactions according to each user's beliefs and preferences. In addition, the current model assumes that agents communicate through well-defined dialogue moves and that their communication is seamless. In reality, however, communication might be affected by factors such as miscommunication or uncertainty. Finally, the current framework is limited to propositional logic, which

Dialogue Type	Initial State	Goal State
Dialectical reconciliation	$KB_i \not\models \varphi$ (or $KB_i \models \neg\varphi$) and $KB_j \models \varphi$	$KB_i \models \varphi$
Persuasion	$KB_i \models \varphi$ and $KB_j \models \neg\varphi$	$KB_j \models \varphi$
Information-seeking	$KB_j \models \varphi$ and $KB_i \not\models \varphi$	$KB_i \cup \varphi \not\models \bot$
Inquiry	$KB_i \not\models \varphi$ and $KB_j \not\models \varphi$	$KB_i \cup KB_j \models \varphi$

Table 4: Dialectical reconciliation contrasted against persuasion, information-seeking, and inquiry dialogue types. Each dialogue type considers a topic (φ) and two agents, the initiating agent (KB_i) and the participating agent (KB_j).

may not be sufficient to express complex relationships and dependencies in real-world domains. To address the limitations and improve the DR-HAI framework, we suggest the following directions for future work.

First, an adaptive approach that tailors arguments to individual users' needs and preferences could further enhance the effectiveness of dialectical reconciliation. Future work could explore methods for personalizing arguments based on user feedback and prior interactions.

Moreover, an interesting direction would be the integration of the DR-HAI framework with user interfaces, such as natural language processing systems, to make the interactions more natural and accessible, especially to non-expert users. This would involve developing techniques for translating the formal arguments generated by the framework into more intuitive, natural language sentences.

Finally, although we used propositional logic to present DR-HAI and demonstrate its efficacy, extending to more expressive logics, such as first-order logic, description logics, or modal logics, will allow for more complex reasoning and argument generation. This would enable the framework to handle a wider range of real-world problems.

References

Besnard, P.; Grégoire, É.; Piette, C.; and Raddaoui, B. 2010. MUS-based generation of arguments and counter-arguments. In *2010 IEEE International Conference on Information Reuse & Integration*, 239–244.

Besnard, P.; and Hunter, A. 2001. A logic-based theory of deductive arguments. *Artificial Intelligence*, 203–235.

Black, E.; and Hunter, A. 2009. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, 19: 173–209.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2020. The Emerging Landscape of Explainable Automated Planning & Decision Making. In *IJCAI*, 4803–4811.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*, 156–163.

Craik, K. J. W. 1967. *The nature of explanation*, volume 445. CUP Archive.

Fan, X.; and Toni, F. 2012. Agent Strategies for ABA-based Information-seeking and Inquiry Dialogues. In *ECAI*, 324–329.

Gordon, T. F. 1994. An inquiry dialogue system. *Artificial Intelligence and Law*, 2: 239–292.

Hamblin, C. L. 1970. *Fallacies*. Methuen and Co Ltd.

Hamblin, C. L. 1971. Mathematical models of dialogue 1. *Theoria*, 37(2): 130–155.

Hitchcock, D.; and Hitchcock, D. 2017. Some principles of rational mutual inquiry. *On Reasoning and Argument: Essays in Informal Logic and on Critical Thinking*, 313–321.

Ignatiev, A.; Morgado, A.; and Marques-Silva, J. 2018. PySAT: A Python Toolkit for Prototyping with SAT Oracles. In *SAT*, 428–437.

Johnson-Laird, P. N. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6.

Kumar, A.; Vasileiou, S. L.; Bancilhon, M.; Ottley, A.; and Yeoh, W. 2022. VizXP: A Visualization Framework for Conveying Explanations to Users in Model Reconciliation Problems. In *Proceedings of the International Conference on Automated Planning and Scheduling*.

Marques-Silva, J.; Heras, F.; Janota, M.; Previti, A.; and Belov, A. 2013. On Computing Minimal Correction Subsets. In *IJCAI*, 615–622.

Mercier, H.; and Sperber, D. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 57–74.

Parsons, S.; Wooldridge, M.; and Amgoud, L. 2003. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3): 347–376.

Plotkin, G. D. 1981. *A structural approach to operational semantics*. Aarhus university.

Prakken, H. 2006. Formal systems for persuasion dialogue. *The knowledge engineering review*, 21(2): 163–188.

Son, T. C.; Nguyen, V.; Vasileiou, S. L.; and Yeoh, W. 2021. Model reconciliation in logic programs. In *European Conference on Logics in Artificial Intelligence*, 393–406.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation. In *ICAPS*, 518–526.

Sreedharan, S.; Hernandez, A. O.; Mishra, A. P.; and Kambhampati, S. 2019. Model-Free Model Reconciliation. In *IJCAI*, 587–594.

Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2022. Explainable Human–AI Interaction: A Planning Perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 16(1): 1–184.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2021. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artificial Intelligence*.

Vasileiou, S. L.; Previti, A.; and Yeoh, W. 2021. On Exploiting Hitting Sets for Model Reconciliation. In *AAAI*.

Vasileiou, S. L.; Yeoh, W.; Son, T. C.; Kumar, A.; Cashmore, M.; and Magazzeni, D. 2022. A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems. *Journal of Artificial Intelligence Research*, 73: 1473–1534.

Walton, D.; and Krabbe, E. C. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.