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

Stylianos Loukas Vasileiou, Ashwin Kumar, William Yeoh

Washington University in St. Louis
{vstylianos, ashwinkumar, wyeoh}@wustl.edu

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 is the central idea behind the field of *explainable AI planning* (XAIP), which aims 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 in XAIP is the human-AI interaction, where the AI agent is expected to communicate effectively with humans to enhance their understanding and address their concerns. Most research efforts within XAIP have been placed on (sequential) decision-making tasks 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 XAIP from different perspectives (Chakraborti, Sreedharan, and Kambhampati 2020), a process called *model reconciliation* has garnered growing interest (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, and the need for explainability arises due to some knowledge asymmetry between these two models that makes the agent's decisions 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 decisions become explicable to the human user.

Despite the growing popularity of model reconciliation approaches, we identify two limitations: (1) It is commonly 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 may lead to incorrect assumptions about the human user's knowledge and preferences, and consequently to unsatisfactory explanations. (2) 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 may fail to work for more complex decisions and tasks that require a more detailed understanding from the human user, especially when there is substantial knowledge discrepancy between the agent and user models.

Our proposal in this paper addresses these two limitations by proposing a new model reconciliation framework that is formulated around *a multi-shot reconciliation paradigm with no assumptions about an a-priori human user model*. Particularly, we introduce the notion of *dialectical reconciliation*,[1] and propose a framework aimed at enhancing human-AI interaction by helping the human user understand the agent's decisions, and, ultimately, the agent's behavior from the perspective of the agent. Note that the goal is not to convince the user to agree and accept the decisions of the agent, as the user may end up disagreeing with the decisions despite understanding why the agent made those decisions. We believe that these kinds of interactions are likely to grow as AI agents play more pervasive roles in our daily life, and we seek to understand the reasons for its behavior.

Our framework, called *Dialectical Reconciliation in Human-AI Interactions* (DR-HAI), is formalized in a game-

[1] Dialectic refers to a discourse between individuals with differing or conflicting views to determine truth via logical argumentation.

theoretic manner, wherein an explainee agent (e.g., a human user) and an explainer agent (e.g., an AI agent) make moves in a game governed by rules that define the appropriateness of their utterances. We describe the *operational semantics* of DR-HAI, that is, 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. Finally, we discuss the concept of explainee understanding in the context of human-AI interactions, and present a simple method for approximating it.

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

> *DR-HAI enables effective human-AI interactions for tasks involving substantial knowledge discrepancies between the explainee and explainer by improving the explainee's understanding of the explainer's decisions and behavior through dialectical reconciliation.*

In the next section, we focus on the role of dialectical reconciliation in fostering human-AI interaction and understanding as well as report some empirical findings from a human-user study.

2 Study: Single-Shot vs. Dialectical Reconciliation

We conducted a study involving a simulated scenario where a human user is presented with the task of troubleshooting an AI home assistant robot named "Roomie" that appears to be disconnected from the internet. In this scenario, the user is given a set of prompts to help them diagnose the problem, 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.

However, the situation is not as straightforward as it seems, and the user is faced with several complications that hinder their ability to resolve the issue. These include an outdated mobile app, an expired license for the wired connection, and a low battery indicated by the flashing light. These obstacles create a realistic scenario for the user to navigate, as they must interact with Roomie to understand the underlying issues in order to get it up and running again.

Overall, this study provides a valuable opportunity to explore how humans interact with AI systems in real-world situations, and how they approach troubleshooting and problem-solving when faced with unexpected obstacles. From a technical standpoint, this narrative allowed us to approximate a human model, facilitating the use of a single-shot model reconciliation-based method as a baseline.

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 "Why are you disconnected?", participants were divided into two groups:

- **Single-Shot (SSR)**: Group 1 received a single-shot model reconciliation explanation, where the human model was assumed to include the information provided during the scenario's introduction. The explanations were computed using a state-of-the-art solver (Vasileiou, Previti, and Yeoh 2021).[2]
- **DR-HAI**: Group 2 interacted with DR-HAI's 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. Finally, the study's hypothesis is:

> **H:** The DR-HAI group will achieve higher comprehension and satisfaction scores compared to the SSR group.

	SSR	**DR-HAI**	**Filtered DR-HAI**
Number of Participants	49	48	37
Comprehension Score (out of 4)	0.30	2.60	**2.94**
Satisfaction Score (out of 5)	2.94	3.56	**3.73**

Table 1: Results of the user study.

Study Results and Discussion: We conducted the user study using Prolific (Palan and Schitter 2018), 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 SSR 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 SSR group, with the filtered group (participants who asked more than one question) achieving the highest scores. This suggests that engagement in dialectical reconciliation, that is, having the ability to ask contrasting questions, enhanced 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 **H**, demonstrating that dialectical reconciliation is more effective in promoting understanding and addressing the human user's concerns compared to single-shot reconciliation.

[2]We used the implementations provided by the authors.

3 Background: Logic-based Argumentation

We now provide a partial review of logic-based argumentation as presented by Besnard and Hunter (Besnard and Hunter 2001). For the sake of brevity, we consider a propositional language $\mathcal{L}$ that utilizes the classical entailment relation, represented by $\models$. We use $\bot$ to denote falsity and assume that a knowledge base KB (a finite set of formulae) is consistent unless specified otherwise.

Our approach relies on an intuitive concept of a logical *argument*, which can be thought of as a set of formulae employed to (classically) prove a particular claim, represented by a formula:

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

We refer to ϕ as the *claim* of the argument and Φ as the *premise* of the argument. The set of all arguments for a claim ϕ from KB is represented by $\mathcal{A}(KB, \phi)$.

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 $A_2 = \langle \{b, g, g \rightarrow a, a \wedge b \rightarrow c\}, c \rangle$*

In the context of dialectical reconciliation, conflicting knowledge exists between participating agents. To account for that, we make use of a general definition of a *counterargument*, that is, an argument opposing another argument by emphasizing points of conflict on the premises or claim of the argument. With a slight abuse of notation:

Definition 2 (Counterargument)**.** *Let KB_i and KB_j be two knowledge bases, and let $A_i = \langle \Phi, \phi \rangle$ 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_j is a* counterargument *for A_i iff $\Phi \cup \Psi \models \bot$.*

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, $A_{j1} = \langle \{l, d, l \wedge d \rightarrow \neg b, \}, \neg b \rangle$ and $A_{j2} = \langle \{e, e \rightarrow \neg c\}, \neg c \rangle$ are two* counterarguments *for A_i from KB_j.*

We denote the set of all counterarguments for an argument A from KB with $\mathcal{C}(KB, A)$.

This overall logic-based argumentation serves as the underlying machinery of our proposed framework.

4 DR-HAI Framework

We now introduce the *Dialectical Reconciliation in Human-AI Interactions* (DR-HAI) framework. In this framework, two agents engage in a dialogue, with one agent taking on the role of an *explainer* (denoted by index r) and the other an *explainee* (denoted by index e). Recall that the scope of the dialogue is to help the explainee *understand* the decisions made by the explainer *from the perspective of the explainer*. Using our user study scenario as an example, the goal is to enable the human user to understand why Roomie thinks that it is disconnected from the internet, independent of whether Roomie is truly disconnected or not.

We use ϕ to represent an explainer's decision and φ to represent the set of all decisions the explainee seeks to understand. In other words, DR-HAI's primary high-level goal is for the explainee to understand the explainer's reasoning, which leads to φ. Three critical assumptions underlie the DR-HAI framework:

- **Agent knowledge bases:** The explainer is associated with a knowledge base KB_r that encodes its own knowledge of the underlying task. The explainee is associated with knowledge base KB_e that encodes *their approximation of the explainer's knowledge*, which can be $\emptyset$. No agent has a-priori access to the other's knowledge base.
- **Explainee queries:** The explainee has a set of possible queries φ for the explainer, where $KB_e \not\models \phi$ (or $KB_e \models \neg\phi$) and $KB_r \models \phi$ for all $\phi \in \varphi$. The explainee initiates a dialogue with an initial query ϕ_1.
- **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.[4]

With these assumptions in place, the primary goal of DR-HAI can be formulated as follows:

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

A critical aspect of this formulation is effectively enabling $KB_e \models \phi$ during the dialogue between explainee and explainer. At a high level, we aim to find a way to help the explainee transition from a state of not understanding a decision ϕ (i.e., $KB_e \not\models \phi$ or $KB_e \models \neg\phi$) to a state of understanding the decision (i.e., $KB_e \models \phi$). Our thesis in this paper is that a natural way of achieving this transition is through an argumentation-based dialogue that facilitates *dialectical reconciliation*.

In the DR-HAI framework, dialectical reconciliation is the process of resolving inconsistencies, misunderstandings, and gaps in knowledge between explainee and explainer. It relies on the exchange of arguments, conflicts, and other dialogue moves that allow the agents to collaboratively construct a shared understanding of the decisions in question. To successfully achieve dialectical reconciliation, the agents follow certain dialogue protocols that guide their interaction:

- Establish a clear dialogue structure, including the use of *locutions* that define permissible speech acts and turn-taking mechanisms.

[3] The minimality constraint maintains argument relevance by eliminating excess premises and pinpointing specific reasons for inferring a claim, while also preventing negative impacts from superfluous premises.

[4] 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.

- Engage in a cooperative and collaborative manner, with both agents focusing on the shared goal of improving the explainee's understanding.
- Employing argumentation techniques, such as offering counterexamples or pointing out logical inconsistencies, to constructively challenge each other's positions and beliefs.

By adhering to these protocols and engaging in dialectical reconciliation, the explainer can help the explainee iteratively refine their knowledge base, ultimately converging on a shared understanding that enables $KB_e \models \phi$ for all decisions in φ.

4.1 Dialectical Reconciliation Dialogue

We now introduce the dialectical reconciliation dialogue type, drawing upon Hamblin's dialectical games framework (Hamblin 1970; Hamblin). Here, a dialogue is viewed as a game-theoretic interaction, where utterances are treated as moves governed by rules that define their applicability. In this context, moves consist of a set of *locutions*, which determine the types of permissible utterances agents can make. To align with the goals of DR-HAI, we define the following set of locutions:

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

The $query$ locution enables the explainee to ask the explainer for an argument that supports the explainee's query. The $support$ locution allows the explainer to provide such an argument that supports the explainee's query. The $refute$ locution permits both agents to present counterarguments, and the $agree\text{-}to\text{-}disagree$ locution allows both agents to acknowledge each other's utterances when no further queries or counterarguments are possible.

We opt for an $agree\text{-}to\text{-}disagree$ locution instead of a simple $agree$ (or $accept$) locution as the goal of DR-HAI is not to convince the explainee about φ but to help them understand φ. An $agree\text{-}to\text{-}disagree$ locution reflects this flexibility, where agents do not have to agree with each other; they only have to acknowledge each other's utterances and understand each other's perspectives. Furthermore, we impose two restrictions: (1) The $query$ locution is only available to the explainee, and (2) The $support$ locution is only available to the explainer. These restrictions are reasonable given the goal of DR-HAI; future work will explore relaxing them.

Locutions are typically instantiated with specific formulae that make up the range of possible *dialogue moves* m_t:

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

where t is an index indicating the dialogue timestep, $x \in \{e, r\}$ denotes 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 use $\mathcal{M}$ to denote the set of all dialogue moves.

We now define the concept 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 turns making and receiving moves. Formally,

Definition 3 (DR-HAI Dialogue)**.** *A* DR-HAI dialogue D *is a sequence of moves* $[m_1, \ldots, m_{|D|}]$ *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* m_t *per timestep* t*.*

We refer to the explainee queries φ made in the dialogue as the *topic* of the dialogue. We use $\mathcal{D}$ to represent the set of all dialogues.

A DR-HAI dialogue is *terminated* at timestep 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 (Terminated DR-HAI Dialogue)**.** *A DR-HAI dialogue* D *is* terminated *at timestep* t *iff* $m_t = \langle e, agree\text{-}to\text{-}disagree, \emptyset \rangle$ *and* $\nexists t' < t$ *s.t.* D *is terminated at timestep* t'*.*

During the dialogue, the agents essentially decide which dialogue moves to make. An agent may have a specific objective in mind when making its decision, such as 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 a function S_x that takes in its current dialogue D, knowledge base KB_x, next timestep t and returns its next dialogue move. While agent strategies can take several forms (e.g., preference-based, probabilistic), for simplicity, we will consider the agents follow an ordered strategy. That is, $S_e(D, KB_e, t) = [refute, query, agree\text{-}to\text{-}disagree]$ and $S_r(D, KB_r, t) = [support, refute, agree\text{-}to\text{-}disagree]$, where the ordered lists show the priorities of dialogue moves for the explainee and explainer, respectively, at timestep $t > 1$.

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 5 (Well-Formed DR-HAI Dialogue)**.** *A DR-HAI dialogue* D *is* well-formed *iff it is terminated at timestep* t *and, for all timesteps* $1 < t' < t$*,* $S_x(D', KB_x, t') = m_{t'}$ *for each move* $m_{t'}$ *from agent* x*, where* $D' \subseteq D$ *consists of the first* $|D'| = t' - 1$ *moves from* D*.*

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 Semantics of DR-HAI Dialogues

At large, not all dialogue moves are valid during a dialogue, as 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*. Fundamentally, a dialogue protocol specifies the *operational semantics* of the dialogue by outlining preconditions and effects for each locution (Plotkin 1981). This means that locutions have action-

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

Table 2: DR-HAI dialogue protocol; when CS_e^T or CS_r^T are used in a condition, it implies that the condition holds for all $1 \leq T \leq t-1$.

like properties, enabling them to change the state of the dialogue.

Table 2 presents how valid dialogue moves m_t ($t > 1$) can be generated in accordance with the DR-HAI dialogue protocol. 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. A subsequent $query$ locution is instantiated with formula Φ if and only if Φ has been uttered by the explainer and not previously queried (preconditions (1)-(2)), and either KB_e 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 the explainee queried Φ in the previous timestep (precondition (1)) and there exists an argument for Φ from KB_r that has not been previously uttered (precondition (2)). The $refute$ locution is instantiated with Φ if and only if Φ has been asserted by the explainer (or explainee) agent at any timestep in the dialogue (precondition (1)) and there exists a counterargument A for Φ from KB_e (or KB_r) that has not been asserted before (precondition (2)). The $agree\text{-}to\text{-}disagree$ locution is uttered if and only if $query$ (or $support$) and $refute$ cannot be uttered by the explainee (or explainer). Finally, the respective agents' commitment stores are updated after each move.

Illustrative Example Consider the following explainer and explainee knowledge bases:

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

Additionally, assume the explainee wants to understand decision c, where $KB_r \models c$ and $KB_e \models \neg c$.

A dialectical reconciliation dialogue is shown below. The dialogue begins with the explainee asking the explainer about c (m_1). The explainer then provides an argument supporting c (m_2). The explainee counters by refuting c (m_3), and the explainer refutes e in the explainee's argument (m_4). Next, the explainee poses a query about h (m_5), and the explainer provides an argument supporting h (m_6). The explainee subsequently refutes f (m_7). Finally, both agents express their agreements to disagree (m_8 and m_9), which leads to the termination of the dialogue.

Dialogue Move	Commitment Store
$m_1 = \langle e, query, c\rangle$	$CS_e^1 = query(c)$
$m_2 = \langle r, support, c\rangle$	$CS_r^2 = \langle\{a, b, a \wedge b \rightarrow c\}, c\rangle$
$m_3 = \langle e, refute, c\rangle$	$CS_e^3 = \langle\{e, e \rightarrow \neg c\}, \neg c\rangle$
$m_4 = \langle r, refute, e\rangle$	$CS_r^4 = \langle\{h, h \rightarrow \neg e\}, \neg e\rangle$
$m_5 = \langle e, query, h\rangle$	$CS_e^5 = query(h)$
$m_6 = \langle r, support, h\rangle$	$CS_r^6 = \langle\{f, f \rightarrow h\}, h\rangle$
$m_7 = \langle e, refute, f\rangle$	$CS_e^7 = \langle\{i, i \rightarrow \neg f\}, \neg f\rangle$
$m_8 = \langle r, agree\text{-}to\text{-}disagree, \emptyset\rangle$	$CS_r^8 = agree\text{-}to\text{-}disagree$
$m_9 = \langle e, agree\text{-}to\text{-}disagree, \emptyset\rangle$	$CS_e^9 = agree\text{-}to\text{-}disagree$

4.3 Properties of DR-HAI Dialogues

In this section, we discuss two important properties for evaluating the quality of a DR-HAI dialogue: *Termination* and *success*.

Termination implies that the dialogue does not continue indefinitely and that it is free from deadlocks, meaning that an agent always has a move to make at any stage of the dialogue.

Theorem 1. *A DR-HAI dialogue always terminates.*

PROOF (SKETCH). First, the operational semantics (see Table 2) outline the constraints and conditions under which each dialogue move can be executed. Second, the agents' knowledge bases are finite, meaning that there are only a limited number of different moves that can be generated, and the agents cannot repeat these moves. As such, the dialogue will not continue indefinitely.

We now prove that a deadlock cannot happen through contradiction. Assume that a deadlock happened, where an agent x does not have any available moves to make and the dialogue has not terminated. There are the following two cases:

- Agent x is an explainee. When the explainee cannot make any $query$ or $refute$ moves, it can always make the

$agree\text{-}to\text{-}disagree$ move since its preconditions are that the preconditions of the $query$ or $refute$ moves do not hold.

- Agent x is an explainer. When the explainer cannot make any $support$ or $refute$ moves, it can always make the $agree\text{-}to\text{-}disagree$ move since its preconditions are that the preconditions of the $support$ or $refute$ moves do not hold.

This contradicts our assumption and the dialogue is thus deadlock-free. Therefore, a DR-HAI dialogue is guaranteed to terminate. □

Now, we want to examine whether the terminated dialogue is successful, that is, if the *goal* of the dialogue is achieved. Recall that the primary goal of DR-HAI is for the explainee agent to understand, from the perspective of the explainer agent, the topic φ of the dialogue, which we formalized as $KB_e \models \phi$ for each $\phi \in \varphi$. For simplicity, we will use $KB_e \models \varphi$ to denote entailment for all $\phi \in \varphi$. This can be easily accomplished by performing a *knowledge update* on KB_e with the explainer's arguments presented in the dialogue. To do this, we adopt a straightforward definition for updating knowledge bases from the literature (Vasileiou et al. 2022):

Definition 6 (Updated Knowledge Base)**.** *An* updated knowledge base KB_e *with argument* A *is* $\widehat{KB_e}^{A} = (KB \cup p(A)) \setminus \gamma$*, where* $p(A)$ *is the premise of argument* A *and* $\gamma \subseteq KB_e \setminus p(A)$*.*

Definition 6 adds the premises of an argument A into KB_e,[5] and retracts a minimal set of formulae γ from KB_e if and only if KB_e is inconsistent with the premises added.

Note that the knowledge base update can be performed by the explainee throughout the dialogue (e.g., when the explainee cannot refute any of the arguments of the explainer). However, for simplicity, 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 necessary to update KB_e in order for the updated KB_e to entail φ. The update could be done sequentially, starting with the latest argument presented by the explainer, and continuing 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 guarantees that $KB_e \models \varphi$ will be satisfied. Therefore, if the goal of the DR-HAI dialogue is met, then we say that the DR-HAI dialogue is *successful*.

Definition 7 (Successful DR-HAI Dialogue)**.** *A terminated DR-HAI dialogue* D *on topic* φ *is* successful *iff* $\widehat{KB_e}^{A} \models \varphi$ *for some* $A \subseteq CS_r$*, where* KB_e *and* CS_r *are the initial knowledge base and commitment store, respectively, of the explainer agent.*

Lastly, if we combine Definition 7 with the restrictions and assumptions underlying the overall DR-HAI framework, we can see that a terminated DR-HAI dialogue is guaranteed to be successful.

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

PROOF (SKETCH). First, recall that the topic of the dialogue φ must be entailed by the explainer (i.e., $KB_r \models \varphi$), which means that an argument for φ from KB_r always exists (Definition 1).

Now, notice that for a terminated dialogue D, the explainer's commitment store CS_r contains the explainer's set of arguments that have been presented during the dialogue. Since $KB_r \models \varphi$, and the arguments in CS_r are derived from KB_r, it follows that using the arguments in CS_r to update the explainee's knowledge base KB_e (w.r.t. Definition 6) will enable $KB_e \models \varphi$, as in the worst case, the entire CS_r will be used to update KB_e.

Therefore, it must be the case that the explainee's knowledge base will eventually entail φ (i.e., $KB_e \models \varphi$) and, as such, a terminated DR-HAI dialogue on topic φ is always successful. □

5 On Explanation and Understanding

Dialectic refers to a method of discourse between individuals with differing views, aimed at arriving at the truth through logical argumentation (van Eemeren et al. 2020). It involves a process of thesis, antithesis, and synthesis, where conflicting ideas are confronted and reconciled to form a new understanding. Drawing inspiration from the dialectic approach, our proposed framework fosters a dialogue between the explainer and the explainee, ultimately aiming to enhance the explainee's understanding of the explainer's decisions.

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.

Measuring the explainee's level of understanding can be a challenging task, as understanding is an abstract concept that may involve various factors, such the explainee's cognitive abilities. From a psychological perspective, nonetheless, 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 combined 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.

[5] Recall that the premise of an argument $A = \langle \Phi, \phi \rangle$ is the first element of the tuple (see Definition 1).

Since our framework is aimed at enhancing the explainee'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 bases of the explainer KB_r and the explainee KB_e. Recall from Section 4 that KB_e is the explainee's approximation of the knowledge base of the explainer. Specifically, *we posit that the explainee's understanding of the explainer's decisions and behavior is likely to improve as the similarities between KB_e and KB_r increase*. Thus, 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 *syntactically* or *semantically*. 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, that is, the logical entailments that result from the knowledge bases. To define a combined syntactic and semantic similarity metric, we employ a weighted Sørensen-Dice similarity index (Dice 1945; Sorensen 1948) as follows:

$$\Sigma = \alpha \cdot \frac{2 \cdot |KB_e \cap KB_r|}{|KB_e| + |KB_r|} + (1-\alpha) \cdot \frac{2 \cdot |E_e \cap E_r|}{|E_e| + |E_r|} \quad (3)$$

where $\alpha \in [0,1]$ is parameter indicating the weight of each metric component, 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 KB_e and KB_r.

Example 3. *Using the example from Section 4.2, 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 by the explainee). The table below shows how the similarity between the knowledge bases evolves with each update:*

#	Premise to Add	Updated KB_e	Similarity Metric
1	$\{f, f \to h\}$	$\{e, e \to \neg c, i, f, f \to h\}$	$\Sigma = 0.5 \cdot \frac{2 \cdot 2}{12} + 0.5 \cdot \frac{2 \cdot 2}{12} = 0.33$
2	$\{h, h \to \neg e\}$	$\{e \to \neg c, i, f, f \to h, h, h \to \neg e\}$	$\Sigma = 0.5 \cdot \frac{2 \cdot 4}{13} + 0.5 \cdot \frac{2 \cdot 3}{11} = 0.58$
3	$\{a, b, a \wedge b \to c\}$	$\{e \to \neg c, i, f, f \to h, h, h \to \neg e, a, b, a \wedge b \to c\}$	$\Sigma = 0.5 \cdot \frac{2 \cdot 7}{16} + 0.5 \cdot \frac{26}{14} = 0.86$

We now consider the similarity measure when using a single-shot model reconciliation explanation on the same example. The single-shot explanation is $\epsilon = \langle \epsilon^+, \epsilon^- \rangle = \langle \{a, b, a \wedge b \to c\}, \{e\} \rangle$ (Vasileiou et al. 2022). Updating KB_e with ϵ, we get $KB_e = (KB_e \cup \epsilon^+) \setminus \epsilon^- = \{a, b, a \wedge b \to c, e \to \neg c, i, i \to \neg f\}$. Computing the similarity of the updated KB_e with KB_r, we get $\Sigma = 0.48$.

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, defined as the computational efforts required to generate arguments and counterarguments, assuming that communication cost is 0.
- **Number of Updates N:** The total count of updates to the explainee's knowledge base after the dialogue, reflecting the volume of new information incorporated by the explainee.
- **Change in Similarity $\Delta\Sigma$:** The change in the similarity between KB_e and KB_r (for $\alpha = 0.5$), comparing their initial (pre-interaction) and final (post-interaction) levels. As we are using two forms of interaction – our proposed *dialectical* reconciliation and a baseline *single-shot* reconciliation (Vasileiou, Previti, and Yeoh 2021), we use $\Delta\Sigma_{DR}$ and $\Delta\Sigma_{SSR}$ to differentiate the two.

Experimental Setup: We created 12 unique pairs of KB_r and KB_e with sizes of $10^2 - 10^4$ by doing the following: (1) We generated random inconsistent propositional KBs of varying sizes of $10^2 - 10^4$; (2) We constructed KB_r by removing a minimal correction set (MCS) from the inconsistent KB to make them consistent;[6] (3) To create KB_e, we controlled the fraction of conflicts between the explainer and explainee with $c = |KB_e|/|KB_r|$. 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. (4) 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 standard method from the literature (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 to simulate a dialogue that resolved conflicts and improved the explainee's understanding.

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.

Experimental Results: Table 3 presents the evaluation results of DR-HAI on various knowledge base sizes $|KB|$ and

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

$\|KB\|$	$c = 0.2$					$c = 0.4$					$c = 0.6$					$c = 0.8$				
	T	L	N	$\Delta\Sigma_{DR}$	$\Delta\Sigma_{SSR}$	T	L	N	$\Delta\Sigma_{DR}$	$\Delta\Sigma_{SSR}$	T	L	N	$\Delta\Sigma_{DR}$	$\Delta\Sigma_{SSR}$	T	L	N	$\Delta\Sigma_{DR}$	$\Delta\Sigma_{SSR}$
2×10^2	0.05s	21	5	11.50%	9.00%	0.04s	11	1	10.10%	9.20%	0.02s	9	2	9.90%	9.20%	0.05s	9	2	9.95%	9.10%
4×10^2	0.07s	15	6	4.50%	2.50%	0.07s	15	6	5.20%	4.76%	0.05s	11	5	5.63%	4.19%	0.06s	11	5	5.60%	5.30%
6×10^2	0.10s	11	5	2.83%	1.37%	0.10s	11	5	2.15%	1.43%	0.20s	23	11	4.27%	1.58%	0.40s	59	29	11.57%	1.92%
8×10^2	0.30s	41	16	5.09%	0.80%	0.40s	43	20	6.45%	0.74%	0.40s	43	9	3.47%	0.73%	0.50s	43	8	3.50%	0.72%
2×10^3	0.50s	5	2	0.53%	0.83%	1.00s	23	9	2.50%	0.50%	2.40s	69	31	5.48%	0.45%	1.10s	25	10	3.57%	0.72%
4×10^3	4.30s	61	29	4.88%	0.37%	5.50s	71	34	6.05%	1.43%	10.20s	109	54	6.72%	0.59%	8.50s	85	42	6.37%	1.73%
6×10^3	3.50s	13	6	0.89%	0.20%	113.00s	87	40	4.65%	0.18%	3.70s	13	6	3.03%	0.24%	8.30s	57	28	4.93%	0.23%
8×10^3	7.60s	43	21	3.30%	1.53%	5.70s	19	9	4.03%	2.86%	37.90s	43	21	5.13%	4.18%	5.60s	19	9	4.45%	4.19%
2×10^4	21.20s	9	4	0.88%	0.15%	21.70s	9	4	0.10%	0.75%	21.60s	9	4	2.25%	0.68%	21.70s	9	4	2.49%	0.07%
4×10^4	38.40s	44	17	3.20%	1.95%	45.50s	66	18	4.30%	2.13%	50.20s	61	16	5.40%	4.19%	55.80s	68	23	6.20%	3.32%
6×10^4	125.30s	90	33	9.40%	7.31%	133.00s	111	52	29.40%	5.15%	129.60s	101	48	33.20%	17.20%	141.50s	120	61	44.90%	21.32%
8×10^4	149.00s	95	32	15.60%	4.79%	155.00s	129	59	25.40%	13.41%	161.50s	121	42	30.10%	21.29%	172.50s	155	72	39.30%	19.47%

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

fractions of conflicts c, allowing us to observe how they influence the dialogue time T, dialogue length L, number of updates N, the change in similarity with DR-HAI $\Delta\Sigma_{DR}$, and the change in similarity with a state-of-the-art single-shot reconciliation approach $\Delta\Sigma_{SSR}$ (Vasileiou, Previti, and Yeoh 2021). The results reveal several trends and insights into the performance of DR-HAI:

- As $|KB|$ increases, T tends to increase as well, indicating that the computational resources required for generating arguments and counterarguments increase with the size of the knowledge base.
- As $|KB|$ and c increase, L and N typically increase as well, suggesting that more dialogue moves and knowledge base updates are needed to resolve inconsistencies as the size of knowledge base and proportion of conflicting formulae grow.
- As N increases, $\Delta\Sigma_{DR}$ often increases as well. This is expected since with each update in the explainee's knowledge base new formulae from the explainer's knowledge base are added, which consequently, increases the explainee's understanding. Moreover, we observe that $\Delta\Sigma_{SSR}$ is generally lower than $\Delta\Sigma_{DR}$, highlighting a benefit of DR-HAI, which is that it allows for a more in-depth conflict resolution approach compared to single-shot explanations.

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.

7 Discussion, Conclusions, and Future Directions

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: *dialectical reconciliation*. Related dialogue types include the following: 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 and comparison of the four dialogue types.

Dialogue Type	**Initial State**	**Goal State**	**Aim**
Dialectical reconciliation	$KB_i \not\models \varphi$ **or** $KB_i \models \neg\varphi$, $KB_j \models \varphi$	$KB_i \models \varphi$	Agent i understands φ.
Persuasion	$KB_i \models \varphi$, $KB_j \models \neg\varphi$	$KB_j \models \varphi$	Agent j is persuaded about φ.
Information-seeking	$KB_j \models \varphi$, $KB_i \not\models \varphi$	$KB_i \cup \varphi \not\models \bot$	Agent i gain knowledge of φ.
Inquiry	$KB_i \not\models \varphi$, $KB_j \not\models \varphi$	$KB_i \cup KB_j \models \varphi$	Agents i and j jointly find a proof for φ.

Table 4: Comparison of the four dialogue types on a topic φ and two agent KBs KB_i (initiating) and KB_j (participating).

Our primary focus is on human-aware AI, particularly model reconciliation problems (MRP) (Sreedharan, Kulkarni, and Kambhampati 2022; Chakraborti et al. 2017; Sreedharan, Srivastava, and Kambhampati 2021; Son et al. 2021; Vasileiou, Previti, and Yeoh 2021; Kumar et al. 2022), following the logic-based approach by Vasileiou et al. (Vasileiou, Previti, and Yeoh 2021; Vasileiou et al. 2022). Our framework addresses two MRP limitations: (1) the explainer agent's assumed knowledge of the human model and (2) single-shot interactions. Notably, Dung et al. (Dung and Son 2022) tackle these limitations using answer set programming, but their approach differs in principle and lacks experimental validation. Our dialectical reconciliation method offers a more interactive, experimentally grounded solution, fostering comprehensive understanding for the ex-

plainee and promoting a collaborative, dynamic explanation process.

Given recent advancements, it is possible for large language models (LLMs) (Bommasani et al. 2021) to solve explainability and reconciliation problems as well. LLMs are remarkable as few-shot learners, skillfully generating well-structured sentences (Brown et al. 2020; Lu et al. 2022). However, their key limitations in providing a solid foundation for logical reasoning, mainly due to their reliance on statistical features for inference, have been well-documented (Rae et al. 2021; Creswell, Shanahan, and Higgins 2023). In contrast, the symbolic nature of DR-HAI inherently provides key theoretical guarantees (e.g., the explanations provided are logically consistent and correct). The ability to carry out multi-step, logically consistent reasoning is essential to engender trust between human users and AI systems.

Limitations and Future Directions: Despite the promising aspects of DR-HAI, it is important to acknowledge its limitations and potential areas for improvement. DR-HAI 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 both 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 may not be sufficient to express complex relationships and dependencies in real-world domains.

To address the limitations and improve the framework, we suggest the following future directions: (1) An adaptive approach that tailors arguments to individual users' needs and preferences could further enhance the effectiveness of dialectical reconciliation. (2) Integration of the DR-HAI framework with user interfaces and natural language processing systems, such as LLMs, to make the interactions more natural and accessible, especially to non-expert users. (3) 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. IEEE.

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.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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.

Creswell, A.; Shanahan, M.; and Higgins, I. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *ICLR*.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302.

Dung, H. T.; and Son, T. C. 2022. On Model Reconciliation: How to Reconcile When Robot Does not Know Human's Model? In Lierler, Y.; Morales, J. F.; Dodaro, C.; Dahl, V.; Gebser, M.; and Tekle, T., eds., *Proceedings 38th International Conference on Logic Programming, ICLP 2022 Technical Communications / Doctoral Consortium, Haifa, Israel, 31st July 2022 - 6th August 2022*, volume 364 of *EPTCS*, 27–48.

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. ???? Mathematical models of dialogue 1. *Theoria*, 37(2).

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

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*.

Lu, K.; Grover, A.; Abbeel, P.; and Mordatch, I. 2022. Pretrained transformers as universal computation engines. In *AAAI*, 7628–7636.

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.

Palan, S.; and Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17: 22–27.

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.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

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.

Sorensen, T. A. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.*, 5: 1–34.

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*.

van Eemeren, F. H.; Garssen, B.; Verheij, B.; Krabbe, E. C. W.; Snoeck Henkemans, A. F.; and Wagemans, J. H. M. 2020. *Handbook of Argumentation Theory*. Springer Dordrecht.

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.