Approximating Human Models During Argumentation-based Dialogues

Yinxu Tang, Stylianos Loukas Vasileiou, William Yeoh

Washington University in St. Louis
{t.yinxu,v.stylianos, wyeoh}@wustl.edu

Abstract

Explainable AI Planning (XAIP) aims to develop AI agents that can effectively explain their decisions and actions to human users, fostering trust and facilitating human-AI collaboration. A key challenge in XAIP is model reconciliation, which seeks to align the mental models of AI agents and humans. While existing approaches often assume a known and deterministic human model, this simplification may not capture the complexities and uncertainties of real-world interactions. In this paper, we propose a novel framework that enables AI agents to learn and update a probabilistic human model through argumentation-based dialogues. Our approach incorporates trust-based and certainty-based update mechanisms, allowing the agent to refine its understanding of the human's mental state based on the human's expressed trust in the agent's arguments and certainty in their own arguments. We employ a probability weighting function inspired by prospect theory to capture the relationship between trust and perceived probability, and use a Bayesian approach to update the agent's probability distribution over possible human models. We conduct a human-subject study to empirically evaluate the effectiveness of our approach in an argumentation scenario, demonstrating its ability to capture the dynamics of human belief formation and adaptation.

Introduction

The increasing integration of AI systems into real-world applications has underscored the critical need for transparency and trust in human-AI interactions. In this landscape, Explainable AI Planning (XAIP) has emerged as a pivotal area of focus (Fox, Long, and Magazzeni 2017), propelled by its promise to develop AI agents capable of explaining their decisions and actions in a manner comprehensible to human users. Central to XAIP is the concept of *model reconciliation* (Chakraborti et al. 2017), a process aimed at aligning the mental models of AI agents and human users to facilitate better understanding and communication. These mental models are typically encoded using planning paradigms (Sreedharan, Srivastava, and Kambhampati 2021; Sreedharan, Chakraborti, and Kambhampati 2021) or logical formalisms (Son et al. 2021; Vasileiou, Previti, and Yeoh 2021; Vasileiou et al. 2022; Vasileiou and Yeoh 2023).

However, a common assumption in most XAIP-related work has been that the AI agent possesses a known and deterministic model of the human user, which is used in the agent's deliberative processes. This simplistic approach may fail to capture the intricate complexities of real-world interactions, as humans often hold beliefs with varying degrees of certainty, and their beliefs evolve dynamically over time. Such simplifications can lead to significant misalignments between AI agents and human users, as the agents might base their decisions or explanations on an inaccurate or incomplete understanding of the human's mental model. Consequently, this can result in decreased human trust and engagement with AI agents, undermining the fundamental goals of the human-AI interaction community and hindering the development of human-compatible AI systems (Russell 2019).

To address this challenge, we propose a novel approach that enables AI agents to adapt their decisions and explanations based on a more nuanced and dynamic understanding of human mental states. We relax the assumption of a deterministic human model and instead posit that the AI agent maintains a probabilistic representation of the human's knowledge. This probabilistic human model is learned and updated dynamically through ongoing interactions, allowing the agent to refine its understanding of the human's mental state over time. To facilitate this learning process, we introduce a framework that learns and updates the probabilistic human model through argumentation-based dialogues (Gordon 1994; Prakken 2006; Parsons, Wooldridge, and Amgoud 2003; Rago, Li, and Toni 2023; Vasileiou et al. 2023). Argumentation provides a natural and expressive mechanism for the agent and the human to exchange information, beliefs, and justifications, allowing for a rich and dynamic interaction. Our learning framework incorporates two complementary update mechanisms: a *trust-based update* and a *certainty-based update*.

The trust-based update mechanism allows the AI agent to adjust its probabilistic human model based on the human's expressed trust in the agent's arguments. Specifically, when the agent presents an argument, the human may not fully accept it as true, but rather evaluate it based on their trust in the agent's argument. We capture this notion of trust-based uncertainty using a trust value $\tau(A_i) \in [0, 1]$ associated with each agent argument A_i, where higher values in-

dicate greater trust. To capture the relationship between the human's trust in the agent's argument and the probability of that argument, we employ a probability weighting function inspired by prospect theory (Tversky and Kahneman 1992). This function maps the human's trust $\tau(A_i)$ in the agent's argument to a probability $p(A_i)$ in a way that accounts for the psychological biases humans exhibit when assessing probabilities under uncertainty.

The certainty-based update mechanism focuses on the human's expressed certainty in their own arguments. When the human puts forward an argument, they may express some degree of uncertainty about it, represented by a probability $p(A_j) \in [0, 1]$. This probability reflects the human's confidence in their own argument A_j, based on factors such as their background knowledge, reasoning process, and awareness of potential counterarguments.

To update the (probabilistic) human model, we use a Bayesian update mechanism. Particularly, we update the agent's probability distribution over possible human models based on the uncertainties associated with the arguments exchanged during the dialogue. At each timestep, when an argument is presented by either the agent or the human, we perform a general update on the probability distribution that increases the probability of the models consistent with the argument, weighted by the argument's associated probability, and decreases the probability of the models inconsistent with the argument.

Finally, to assess the framework's ability to approximate human models through argumentation-based dialogues, we conduct a human-user study that simulates a decision-making scenario. Our findings demonstrate the feasbility of our approach to dynamically approximate a human model as a probability distribution, leading to increased trust and satisfaction among participants.

The main contributions of this paper are as follows:

- We propose a novel framework for learning and updating a probabilistic human model through argumentation-based dialogues, where we incorporate trust-based and certainty-based update mechanisms.
- We conduct a human-user study to empirically evaluate the effectiveness of our approach in decision-making scenario, demonstrating its ability to capture the dynamics of human belief formation and adaptation.

Related Work

We situate our work with respect to explainable AI planning and argumentation-based dialogues.

Explainable AI Planning

Explainable AI planning (XAIP) aims to foster trust, facilitate human-AI collaboration, and enable effective decision support in complex domains by providing users with understandable explanations of planning processes and decision-making (Fox, Long, and Magazzeni 2017). XAIP has been applied in various domains, including robotics (Setchi, Dehkordi, and Khan 2020), healthcare (Saraswat et al. 2022), and beyond, highlighting its broad applicability and potential impact.

A central focus of XAIP research has been on the concept of model reconciliation (Chakraborti et al. 2017; Sreedharan, Chakraborti, and Kambhampati 2021; Vasileiou, Previti, and Yeoh 2021; Vasileiou et al. 2022; Vasileiou and Yeoh 2023), which seeks to align the mental models of AI agents and human users. However, the original framework often assumes that the agent has perfect knowledge of the human's model a priori, which can lead to incorrect assumptions and suboptimal explanations. To address this limitation, recent works have focused on relaxing the assumptions made about the human model. Notably, Dung and Son (2022) tackled this limitation from the perspective of answer set programming, tying their approach exclusively to planning problems.

On the other hand, some related work tackled this limitation by considering uncertainty about the human's model. In this context, Sreedharan, Chakraborti, and Kambhampati (2018) propose a framework for reconciling with a set of possible human models, demonstrating how it can be used to provide explanations to multiple human users simultaneously. This work highlights the importance of accounting for the inherent uncertainty and variability in human mental models. In a related study, Sreedharan, Srivastava, and Kambhampati (2018) developed a method for estimating the mental model from a provided foil, further emphasizing the need for techniques that can infer and update the agent's understanding of the human's model based on the available information.

Argumentation-based Dialogues

Argumentation-based dialogues have been developed to aid two (or more) agents in making decisions regarding their goals and plans. In this context, two agents with shared goals will only endorse plans that align with their beliefs. The literature on argumentation-based dialogues spans multiple disciplines, including AI (Bench-Capon and Dunne 2007), legal reasoning (Zhong et al. 2014), and multi-agent systems (Nielsen and Parsons 2006), underlining the broad applicability and interdisciplinary nature of this research area.

Belesiotis, Rovatsos, and Rahwan (2010) propose an abstract argumentation-based protocol that enables two agents to deliberate on their proposals until they reach an agreement, guided by the persuasion-aligned planning beliefs of the agents. This work demonstrates the potential of argumentation-based approaches for facilitating collaborative decision-making and consensus-building in multi-agent settings. Argumentation-based explanation have also gained a lot of traction (Fan and Toni 2015; Shams et al. 2016; Fan 2018; Collins, Magazzeni, and Parsons 2019; **?**; Budán et al. 2020; Dennis and Oren 2022; Rago, Li, and Toni 2023). These works primarily focus on explanations whose justification is provided through argumentation semantics using specific dialogue formalizations, establishing an equivalence between the dialogues and the argumentation semantics. At the intersection of argumentation-based dialogues and XAIP is the work by Vasileiou et al. (2023), where the authors proposed a dialectical reconciliation dialogue between an AI agent and a human user with no assumptions about a known human model. The goal of this dialogue is to improve the un-

derstanding of the human user's understanding of the agent's decisions. While these approaches provide a solid foundation for argumentation-based explanations, they do not explicitly consider the uncertainty inherent in human-agent interactions.

Most closely related to our setting, another line of research investigated probabilistic argumentation ((Hunter 2013, 2014, 2022)) and introduces uncertainties in argumentation. Specifically, uncertainties are represented by probabilistic measures, e.g., probabilities or degrees of belief, and assigned to propositions or arguments. These works build upon probabilistic argument graphs, as defined in (Dung and Thang 2010; Li, Oren, and Norman 2011), providing a formal framework for reasoning about uncertain arguments.

Building upon ideas from XAIP and argumentation-based dialogues, we provide a probabilistic approach to modeling and updating the agent's representation of the human's model. Compared with previous works, our framework enables a more nuanced and adaptive approach to model reconciliation in XAIP by maintaining a probability distribution over possible mental models and updating it based on the human's trust and certainty feedback.

Background

We assume classical propositional logic for describing aspects of the world. We consider a finite (propositional) language $\mathcal{L}$ that utilizes the classical entailment relation, represented by $\models$. The set of *models* (i.e. possible words) of $\mathcal{L}$ is denoted by $\mathcal{M}$, where each model $m_i \in \mathcal{M}$ is an assignment of true or false to the formulae of $\mathcal{L}$ defined in the usual way for classical logic. For $\phi \in \mathcal{L}$, let $\texttt{Mod}(\phi) = \{m_i \in \mathcal{M} \mid m_i \models \phi\}$ denote the set of models of ϕ.

Logic-based Argumentation

We provide a partial review of logic-based argumentation (Besnard and Hunter 2014). Our framework relies on an intuitive understanding of a logical *argument*, which is essentially a set of formulae used to prove a specific claim.

Definition 1 (Argument)**.** *Let $\mathcal{L}$ be the language and $\varphi \in \mathcal{L}$ a formula. An argument for φ is defined as $A = \langle \Phi, \varphi \rangle$ such that: (i) $\Phi \subseteq \mathcal{L}$; (ii) $\Phi \models \varphi$; (iii) $\Phi \not\models \bot$; and (iv) $\nexists \Phi' \subset \Phi$ s.t. $\Phi' \models \varphi$.*

We refer to φ as the *claim* of the argument, and Φ as the *premise* of the argument.

Example 1. *Let $\mathcal{L}$ be a propositional language made up of variables $\{a, b, c, d, e\}$. Then, $A_1 = \langle \{a, b, a \wedge b \rightarrow c\}, c \rangle$ and $A_2 = \langle \{b, d, d \rightarrow a, a \wedge b \rightarrow c\}, c \rangle$ are two arguments for c.*

We incorporate a general definition of a *counterargument* to address conflicting knowledge among agents. A counterargument is defined as an argument that opposes another argument by highlighting conflicts regarding the premises or claims. Specifically,

Definition 2 (Counterargument)**.** *Let $\mathcal{L}$ be the language, and let $A_1 = \langle \Phi, \varphi \rangle$ and $A_2 = \langle \Psi, \psi \rangle$ be two arguments for φ and ψ, respectively. We say that A_2 is a counterargument for A_1 iff $\Phi \cup \Psi \models \bot$.*

Example 2. *Let $\mathcal{L}$ be a propositional language made up of variables $\{a, b, c, d, e\}$, and let $A_1 = \langle \{a, b, a \wedge b \rightarrow c\}, c \rangle$ be an argument for c. Then, $A_2 = \langle \{f, d, f \wedge d \rightarrow \neg b\}, \neg b \rangle$ and $A_3 = \langle \{e, e \rightarrow \neg c\}, \neg c \rangle$ are two counterarguments for A_1.*

Modeling Uncertainty in Propositional Logic

Building on a propositional language $\mathcal{L}$, we can model the uncertainty of arbitrary formulae using a *probability distribution* over the models $\mathcal{M}$ of $\mathcal{L}$. Formally,

Definition 3 (Probability Distribution)**.** *Let $\mathcal{M}$ be the set of models of the language $\mathcal{L}$. A probability distribution P on $\mathcal{M}$ is a function $P : \mathcal{M} \mapsto [0, 1]$ such that $\sum_{m \in \mathcal{M}} P(m) = 1$.*

In essence, a probability distribution over the models of $\mathcal{L}$ creates a *ranking* between those models with respect to how likely they are to be true. This then allows us to quantify the uncertainty in a formula as follows:

Definition 4 (Degree of Belief)**.** *Let $\mathcal{M}$ be the set of models of language $\mathcal{L}$ and P a probability distribution over $\mathcal{M}$. The degree of belief of a formula $\phi \in \mathcal{L}$ is $P(\phi) = \sum_{m \models \phi} P(m)$.*

We may refer to $P(\phi)$ as degree of belief or probability of ϕ interchangeably. Note that this approach to probabilities is essentially equivalent to probabilities assigned directly to the formulae (Bacchus 1990).

Example 3. *Let $\mathcal{L}$ be a propositional language with variables $\{a, b\}$. An example of a probability distribution over the models $\mathcal{M}$ of $\mathcal{L}$ is shown in Table 1.*

	m_1	m_2	m_3	m_4
a	True	True	False	False
b	True	False	True	False
$P(m_i)$	0.1	0.2	0.4	0.3

Table 1: An example of probability distribution over models.

Then, a has degree of belief $P(a) = P(m_1) + P(m_2) = 0.3$. Similarly, $a \rightarrow b$ has degree of belief $P(a \rightarrow b) = P(m_1) + P(m_3) + P(m_4) = 0.8$..

Approximating Human Models During Argumentation-based Dialogues

In this section, we introduce a framework that allows an agent to progressively update its approximation of the human model over the course of an argumentation-based dialogue (Vasileiou et al. 2023).

Problem Setting and Assumptions

We consider an argumentation-based dialogue between two participants: an agent (a) and a human (h). We make the following key assumptions:

- **Shared Domain Language**: Both a and h have access to and communicate in the same propositional language $\mathcal{L}$, with a shared vocabulary of atomic variables. This allows them to construct domain-specific formulae.

- **Uncertain Human Model**: The agent maintains a probabilistic model of the human's knowledge, represented by a probability distribution P_h over the possible models $\mathcal{M}$ of the human's knowledge at each step of the dialogue. This distribution captures the agent's uncertainty about the human's knowledge. Initially, the agent assumes a uniform prior $P_h^{t_0}(m) = \frac{1}{|\mathcal{M}|}$ for all $m \in \mathcal{M}$, representing agnosticism about the human model.
- **Argument Traces**: We assume access to a (finite) argument trace $\mathcal{D} = \langle (A_1, x_1)^{t_1}, (A_2, x_2)^{t_2}, \ldots \rangle$ produced by the dialogue, where each $(A_i, x_i)^{t_i}$ represents an argument A_i put forward by participant $x_i \in \{a, h\}$ at timestep t_i.

Handling Argument Uncertainty

In real-world argumentation, the arguments put forward by both the agent and the human often come with some degree of uncertainty. This uncertainty can arise from various sources, such as incomplete or imprecise knowledge, subjective interpretations, or lack of confidence in the reasoning process. In our framework, we consider two types of uncertainty associated with arguments:

- **Uncertainty in Agent's Arguments**: When the agent presents an argument A_i at timestep t_i, the human may not fully accept the argument as true, but rather evaluate it based on their trust in the agent. Intuitively, if the human has a high level of trust in the agent, they are more likely to assign a high probability to the agent's argument, indicating that they believe it is likely to be true or valid. Conversely, if the human has low trust in the agent, they may assign a lower objective probability to the argument, reflecting their doubts or skepticism about its correctness. We capture this notion of trust-based uncertainty using a trust value $\tau(A_i) \in [0, 1]$ associated with each agent argument A_i, where higher values indicate greater trust.
- **Uncertainty in Human's Arguments**: When the human puts forward an argument A_j at timestep t_j, they may express some degree of uncertainty about it, represented by a probability $p(A_j) \in [0, 1]$. This probability reflects the human's confidence in their own argument, based on factors such as their background knowledge, reasoning process, and awareness of potential counterarguments. Higher values of $p(A_j)$ indicate greater certainty in the argument.

Now, according to prospect theory (Kahneman and Tversi 1979), the probability of an event may not align with the "subjective" perception of that probability, that is people tend to overweight small probabilities and underweight moderate to high probabilities when making decisions under uncertainty (Fox and Poldrack 2009). To this end, we propose a probability weighting function (Gonzalez and Wu 1999) to describe the relationship between the two. In our scenario, we define the *trust value* of argument A_i as the human's subjective perception of the argument's uncertainty, denoted by $\tau(A_i)$. To capture its relationship with the probability of the argument $p(A_i)$, we use the following sigmoidal function (Tversky and Kahneman 1992).

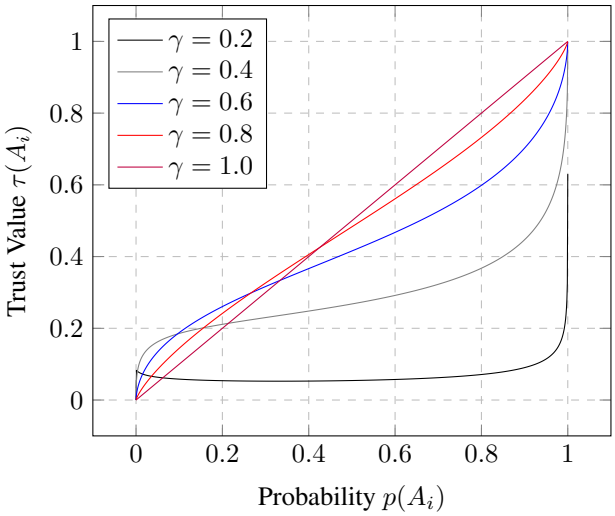

Figure 1: Probability weighting function with $\gamma = 0.2, 0.4, 0.6, 0.8$ and 1.

$$\tau(A_i) = \frac{p(A_i)^\gamma}{(p(A_i)^\gamma + (1 - p(A_i))^\gamma)^{1/\gamma}}, \tag{1}$$

where $\gamma \in (0, 1)$ is a parameter that controls the degree of this nonlinear distortion.
Specifically, lower values of γ (closer to 0) indicate excessive distortion (i.e., overweighting or underweighting) of the objective probability, while higher values (closer to 1) indicate a nearly linear relationship between the trust and the objective probability. The examples of relationships are shown in Figure 1.

The key idea behind using this function in our argumentation model is that it provides a psychologically plausible way to map the human's trust in an argument to an objective probability that is not just a linear scaling of trust. By accounting for the nonlinear biases in human probability assessment, the function allows the agent to more accurately model how the human is likely to respond to arguments of varying degrees of trustworthiness.

Updating the Human Model

We employ a Bayesian approach to update the agent's probability distribution P_h over possible human models based on the uncertainties associated with the arguments exchanged during the dialogue. At each timestep t_i, when an argument A_i is presented by either the agent or the human, we perform the following general update on the probability distribution:

$$P_h^{t_i}(m) = \begin{cases} \dfrac{P_h^{t_{i-1}}(m)}{\sum_{m \models A_i} P_h^{t_{i-1}}(m)} \cdot p(A_i) & \text{if } m \models A_i \\ \dfrac{P_h^{t_{i-1}}(m)}{\sum_{m \not\models A_i} P_h^{t_{i-1}}(m)} \cdot (1 - p(A_i)) & \text{if } m \not\models A_i \end{cases} \tag{2}$$

where $m \models A_i$ denotes that model m is consistent with argument A_i, i.e., the premises and conclusion of A_i hold in m, and $p(A_i)$ is the probability associated with the argument. The proportionality constant is chosen to ensure that $P_h^{t_i}$ remains a valid probability distribution.

Intuitively, the update mechanism in Eq. (2) increases the probability of human models that are consistent with the presented argument, weighted by the argument's associated probability $p(A_i)$. Models that are inconsistent with the argument have their probabilities decreased accordingly. The higher the probability of the argument, the more the distribution shifts towards consistent models.

Note that the probability $p(A_i)$ is determined based on the source of the argument. If A_i is presented by the agent, $p(A_i)$ is the objective probability derived from the human's trust in the argument, $\tau(A_i)$, using the probability weighting function Eq. (1). Specifically, $p(A_i)$ is obtained by numerically inverting Eq. (1) to solve for $p(A_i)$ given $\tau(A_i)$.

Example 4. *Consider a dialogue where at timestep t_1, the agent asserts the argument $A_1 = \langle\{a, a \rightarrow b\}, \{b\}\rangle$. The human assigns a trust value of $\tau(A_1) = 0.6$ to this argument. Assuming $\gamma = 0.85$, the objective probability of A_1 is computed using Eq.* (1)*:*

$$0.6 = \frac{p(A_1)^{0.85}}{[p(A_1)^{0.85} + (1 - p(A_1))^{0.85}]^{\frac{1}{0.85}}}$$

Solving for $p(A_1)$, we get $p(A_1) \approx 0.62$. Suppose there are four possible models, $\mathcal{M} = \{m_1, m_2, m_3, m_4\}$, with a uniform prior distribution $P_h^{t_0}(m_1) = \ldots = P_h^{t_0}(m_4) = 0.25$. Let m_1 be the model that entails the premises of A_1, i.e., $m_1 \models \{a, a \rightarrow b\}$. Applying the update mechanism from Eq. (2)*, we get:*

$$P_h^{t_1}(m_1) = \frac{0.25}{0.25} \cdot 0.62 = 0.62$$
$$P_h^{t_1}(m_k) = \frac{0.25}{0.25 + 0.25 + 0.25} \cdot 0.38 = 0.126 \ (k = 2, 3, 4)$$

After this update, the model m_1 that is consistent with the agent's argument has a higher probability than the other three models, reflecting the human's moderate trust in the argument.

On the other hand, if A_i is an argument presented by the human, $p(A_i)$ is the probability directly expressed by the human for their own argument.

Example 5. *Continuing the previous example, suppose at timestep t_2, the human presents the argument $A_2 = \langle\{\neg a\}, \{\neg a\}\rangle$ with probability $p(A_2) = 0.9$.*

Let m_3 and m_4 be the models that entail the premise of A_2. Applying the update mechanism to the distribution resulting from the previous (objective) probability update, we get:

$$P_h^{t_2}(m_1) = \frac{0.62}{0.62 + 0.126} \cdot 0.1 = 0.083$$
$$P_h^{t_2}(m_2) = \frac{0.126}{0.62 + 0.126} \cdot 0.1 = 0.017$$
$$P_h^{t_2}(m_3) = P_h^{t_2}(m_4) = \frac{0.126}{0.126 + 0.126} \cdot 0.9 = 0.45$$

After this update, the models m_3 and m_4 that are consistent with the human's argument have much higher probability than the models consistent with the agent's previous argument.

By applying this update rule iteratively according to the sequence of arguments in the dialogue trace $\mathcal{D}$, the agent can gradually refine its estimate of the human model distribution P_h. The refined distribution incorporates information about the uncertainties associated with both the agent's and human's arguments, providing a more nuanced and psychologically grounded estimate of the human's knowledge.

It is worth noting that this update rule assumes that the human's trust in the agent's arguments and their own expressed probabilities are well-calibrated and consistent across the dialogue. In practice, there may be situations where the human's probability assessments are inconsistent or biased. Handling such inconsistencies and biases is an important challenge for future work. Nevertheless, the proposed update rule provides a simple and principled way to integrate argument uncertainties into the agent's modeling of the human's mental state, enabling more effective adaptation of the agent's argumentative strategies to the individual human.

Empirical Evaluation: Human-User Study

To evaluate the effectiveness of our proposed framework for approximating human models during argumentation-based dialogues, we conducted a user study simulating a real-world scenario. In this study, participants interacted with an AI assistant named "Blitzcrank" to assess the suitability of a fictional venue, "Luminara Gardens", for hosting a company team-building event. The study aimed to investigate the dynamics of human-AI interaction, the AI's ability to gauge participants' understanding, and changes in participants' trust levels throughout the dialogue.

Based on our proposed framework and the designed scenario, we formulated the following hypotheses:

H_1: Our framework can effectively approximate a probability distribution that captures the participants' knowledge during argumentation-based dialogues.

H_2: Participants' trust in the AI assistant increases as the interaction progresses.

Study Design

Dialogue Design: The study consisted of a series of interaction rounds between each participant and Blitzcrank. In each round, the participants were presented with a set of Blitzcrank's arguments regarding the suitability of Luminara Gardens for the team-building event. The arguments varied in their level of informativeness and persuasiveness, reflecting different degrees of argument strength.

After receiving an argument from Blitzcrank, the participants were asked to select their level of trust in the argument from four options: almost complete trust ($\tau = 0.9$), high trust ($\tau = 0.7$), average trust ($\tau = 0.5$), or low trust ($\tau =$

Trust Level		Almost Complete Trust	High Trust	Average Trust	Low Trust
Trust Value		$\tau = 0.9$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
Objective Probability	$\gamma = 0.4$	1.000	0.990	0.937	0.150
	$\gamma = 0.5$	1.000	0.959	0.804	0.104
	$\gamma = 0.6$	0.989	0.898	0.657	0.114
	$\gamma = 0.7$	0.972	0.826	0.566	0.133
	$\gamma = 0.8$	0.949	0.765	0.522	0.155
	$\gamma = 0.9$	0.922	0.724	0.504	0.178

Table 2: Mapping of trust levels to the probabilities of Blitzcrank's arguments.

Certainty Level	Probability	Linguistic Cues
High Certainty	0.9	"I am confident that..." "I am certain that..." "There is no doubt that..."
Moderate Certainty	0.7	"It seems probable that..." "It's quite likely that..." "There's a good chance that..."
Neutral Uncertainty	0.5	"I'm not entirely sure, but..." "It could be the case that..." "There's a possibility that..."
Moderate Uncertainty	0.3	"There's some doubt as to whether..." "It's questionable whether..." "It's uncertain if..."
High Uncertainty	0.1	"I'm not confident in saying..." "It's hard to say for sure..." "There's significant uncertainty..."

Table 3: Classification of participant arguments by certainty level according to linguistic cues in their arguments.

0.2). These trust levels $\tau(A_i)$ were mapped to objective argument probabilities $p(A_i)$ using the probability weighting function Eq. (1). In our experiments, we computed $p(A_i)$ numerically with the Newton–Raphson method (Kelley 2003). Table 2 shows the computed probabilities with respect to the trust levels. Note that here we select γ from the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Next, the participants were presented with a set of five candidate counterarguments, each associated with a certainty level. The certainty level was inferred from linguistic cues in the arguments (e.g., "I'm quite sure that...", "Is it possible that..."). The participant selected one of these counterarguments to present to Blitzcrank. Table 3 shows the classification of the participants' arguments by certainty level.

This process of argument presentation, trust assessment, and counterargument selection constituted one interaction round. After each round, the participants were asked to rank four different perspectives on Luminara Gardens' suitability for the event, based on their current understanding.[1] These perspectives represented models, and the rankings provided a measure to assess the participant's understanding at each stage. Rank 1 indicated the most likely perspective (model), while rank 4 indicated the least likely one.

The dialogue continued for a fixed number of rounds (up to three) or until the participants chose to end the interaction. The specific arguments, counterarguments, and perspectives used in the study were designed to cover a range of aspects related to Luminara Gardens' suitability, such as venue capacity, catering options, entertainment facilities, and pricing.

[1]Note that the user model would be potentially changing over rounds, depending on the trust level of the agent's argument and the uncertainty level of the user's argument in the subsequent interactions.

Participant Details: We recruited 150 participants via the Prolific platform (Palan and Schitter 2018). Participants were required to be fluent in English and were compensated \$2.00 for their time. Out of the 150 participants, 143 completed the study satisfactorily by passing attention checks and providing coherent responses.

Participants were divided into two groups based on the number of interaction rounds they chose to complete: Group A (44 participants) engaged in two rounds of interaction with Blitzcrank, while Group B (99 participants) engaged in three rounds, which was the maximum allowed. The choice of the number of rounds was up to the participants, reflecting their willingness to engage in a shorter or longer dialogue with the AI assistant.

Post-Study Questionnaire: After completing the dialogue, participants answered a post-study questionnaire containing five Likert-scale items (1 - strongly disagree to 5 - strongly agree). Three items assessed changes in trust levels across the interaction rounds, while the remaining two items evaluated overall satisfaction with the interaction and the quality of Blitzcrank's arguments.

Variants and Baselines

In the following, we introduce three variants of our proposed method:

- *Variant 1: Personalization Upper Bound*: Observe that every participant has a distinct relationship between his/her trust level and the objective probability. As such, we personalize the specific value of γ in Eq. (1), for every individual, that optimizes the distribution of Spearman's rank correlation using all user data in this variation.
- *Variant 2: Personalization I*: (Group A) For each participant in Group A, we use the data from the first interaction process to determine the personal predicted value of γ, which is then applied to the second interaction process; (Group B) For each participant in Group B, we use the data from the first two interaction processes to determine the personal predicted value of γ, which is then applied to the final interaction process.
- *Variant 3: Personalization II*: Unlike Variant 2, for each participant in Group B, we use only the data from the first interaction process to determine the personal predicted value of γ, which is then applied to the last two interaction processes.

To evaluate our framework, we compare our proposed method with the following three baselines:

- *Baseline 1*: The trust update does not apply the weighting function (i.e., $\gamma = 1$ in Eq. (1)). The probability distribution update involves assigning a probability to each model that entails the argument, followed by normalization. For-

Methods	Trust Weighting	Probability Update
Baseline 1	$\gamma = 1$ in Eq. (1)	Eq. (3)
Baseline 2	$\gamma = 1$ in Eq. (1)	Eq. (2)
Baseline 3	Eq. (1)	Eq. (3)
Proposed Method	Eq. (1)	Eq. (2)

Table 4: Baseline methods

mally,

$$P_h^{t_i}(m) = \begin{cases} \frac{p(A_i)}{Z} & \text{if } m \models A_i \\ \frac{P_h^{t_{i-1}}(m)}{Z} & \text{if } m \not\models A_i \end{cases} \quad (3)$$

where $Z = \sum_{m \models A_i} p(A_i) + \sum_{m \not\models A_i} P_h^{t_{i-1}}(m)$.

- *Baseline 2*: The trust update does not apply the weighting function (i.e., $\gamma = 1$ in Eq. (1)). The probability distribution update follows Eq. (2).
- *Baseline 3*: The trust update applies weighting function Eq. (1). The probability distribution update follows Eq. (3).

Table 4 shows the trust weighting rule and probability update of baseline methods.

Evaluation Metrics

To quantitatively evaluate our framework's performance in approximating human models and assess the significance of trust changes, we employed the following metrics:

- **Spearman's Rank Correlation**: We computed Spearman's rank correlation coefficient (Spearman 1904) (ρ) between the participant's perspective rankings and the rankings generated by our framework at each interaction round. A high positive correlation indicates that our framework effectively approximates the participant's understanding of the situation.
- **Student's t-Test**: To determine whether participants' trust levels increased between interaction rounds, we conducted paired t-tests (Student 1908) with p-value 0.05 comparing trust scores across rounds. Separate tests were performed for Group A (comparing trust between rounds 1 and 2) and Group B (comparing trust between rounds 1 and 2, and between rounds 2 and 3).

Results and Discussion

The results of our user study provide strong support for both hypotheses H_1 and H_2.

Figure 2 displays the distribution of Spearman's rank correlation coefficients over different values of γ across all participants. The majority of coefficients are above 0.75, indicating a substantial agreement between the participants' perspective rankings and those generated by our framework. This finding suggests that our approach effectively approximates a probability distribution that captures the participants' knowledge during argumentation-based dialogues,

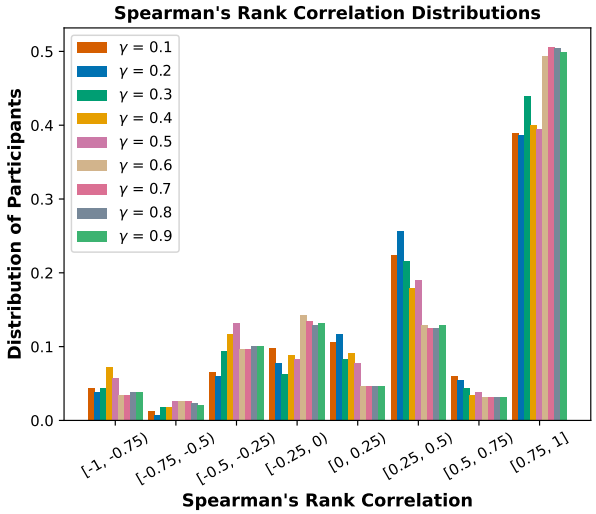

Figure 2: Spearman's rank correlation distributions over different values of γ.

supporting hypothesis H_1. Moreover, $\gamma = 0.7$ produces the best distribution among all chosen values.

Figure 4 illustrates the distribution of Spearman's rank correlation coefficients for different variants. The Personalization Upper Bound achieves the optimal distribution of the proposed method. Moreover, Personalization I yields a better distribution than Personalization II because it utilizes more user data during the prediction process. Specifically, Personalization I can be considered to produce the best achievable distribution of our proposed method, given that obtaining the optimal distribution produced by the Personalization Upper Bound by using all user data for determining individual γ values is unrealistic. The effectiveness of all variants demonstrates the flexibility and potential of our proposed methods, thereby supporting hypothesis H_1.

Figure 4 shows that our proposed method outperforms the baseline methods in the distribution of Spearman's rank correlation coefficients, which hereby supports hypothesis H_1. Specifically, our method enables 55% of correlation values to lie in the region of $[0.75, 1]$, exceeding the $46\%, 50\%, 46\%$ provided by the baseline methods.

Table 5 shows a gradual increase in the average trust score as the dialogue proceeds. Besides, Table 6 presents the results of the t-tests comparing trust scores between interaction rounds. For both Group A and Group B, we observe statistically significant increases in trust scores from round 1 to round 2 ($p_{1,2} < 0.05$). Additionally, for Group B, trust scores significantly increase from round 2 to round 3 ($p_{2,3} < 0.001$). Such results provide compelling evidence for hypothesis H_2, indicating that participants' trust in the AI assistant grows as the dialogue progresses and the assistant provides more relevant and persuasive arguments.

Our user study demonstrates the effectiveness of our proposed framework in an argumentation-based dialogue scenario. The results highlight the framework's ability to dynamically approximate a human model as a probability distribution, leading to increased trust and satisfaction

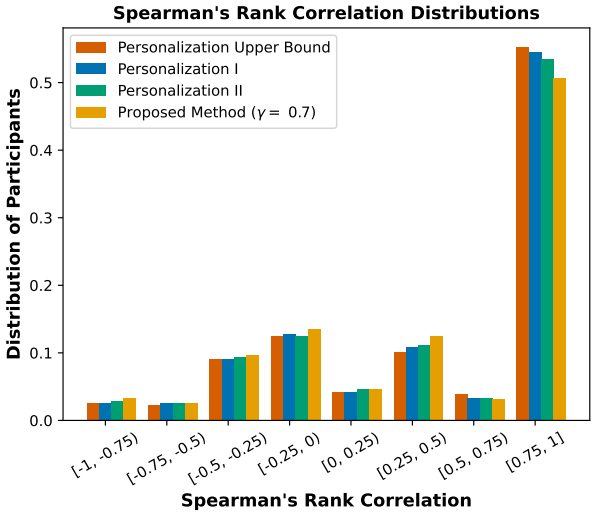

Figure 3: Comparisons of Spearman's rank correlation distributions under different variants.

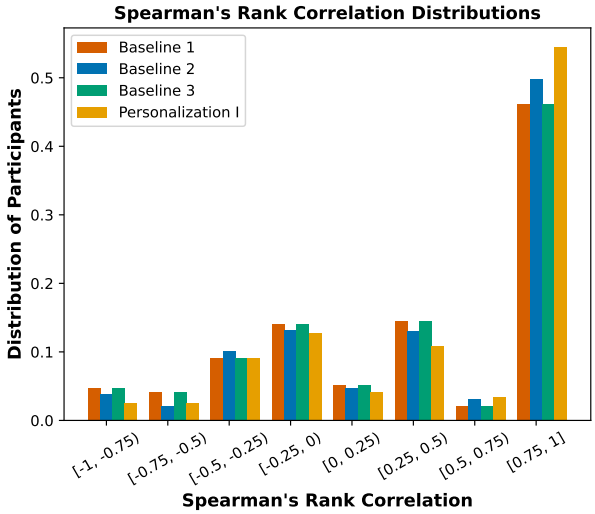

Figure 4: Comparisons of Spearman's rank correlation distributions under different baselines. Note that Baselines 1 and 3 have the same distribution since the chosen values of γ do not significantly influence the outcome due to the probability update function Eq. (3).

Average Trust Score	Group A	Group B
First Round	0.511	0.508
Second Round	0.634	0.552
Third Round	N/A	0.662

Table 5: Average trust degree in different rounds.

	Group A (2 rounds)	Group B (3 rounds)
$p_{1,2}$	0.00011297	0.04303665
$p_{2,3}$	N/A	0.000001278

Table 6: Statistical significance ($p < 0.05$) of trust change.

among participants. Moreover, the post-study questionnaire responses further corroborate these findings, with participants reporting high levels of satisfaction with the interaction and the quality of Blitzcrank's explanations.

Discussion and Conclusions

In this paper, we introduced a novel framework for approximating human mental models during argumentation-based dialogues. Our approach leverages a Bayesian belief update mechanism to refine a probability distribution over possible human models based on the arguments exchanged throughout the dialogue. By incorporating uncertainty estimates for both the agent's and human's arguments, our framework provides a principled way to reason about the human's evolving knowledge state and perspectives.

The results of our human-subject study demonstrate the potential effectiveness of our framework in an applied argumentation setting. The high correlation between the rankings generated by our approach and the participants' actual perspective rankings suggests that our framework can capture some of the dynamics of human belief formation during argumentative interactions.

However, it is important to emphasize that this work is still in its early stages, and further research is needed to validate and refine the proposed framework. Our study had a limited sample size and focused on a single argumentation domain, so the generalizability of the findings to other contexts remains to be established. Moreover, the framework currently makes several simplifying assumptions, such as the consistency and calibration of human probability judgments, which may not hold in real-world settings.

One of the key aspects of our framework is the notion of argument-specific trust, which shapes the human's perception of the agent's arguments. We model trust as being influenced by factors such as perceived relevance, logical strength, consistency with prior beliefs, and clarity. The trust value assigned to each argument is treated as an observable input to the belief update process. However, our current framework does not explicitly model a global notion of trust, i.e., the human's overall trust in the agent across the entire dialogue. Extending the framework to incorporate a global trust component, and investigating its interplay with argument-specific trust, is an important direction for future work.

Another important consideration in practical argumentation systems is how to elicit the human's certainty levels for their own arguments. We envision several possible approaches, including explicit probability input, categorical confidence ratings, and inference from linguistic cues. A combination of these methods, allowing for both system-generated estimates and user adjustments, may provide a good balance of accuracy and usability. However, further empirical studies are needed to understand the impact of different elicitation methods on the quality and calibration of probability estimates in real-world settings.

In our study, the human users were asked to quantify their trust in the agent's argument A_i with a trust value $\tau(A_i) \in$

$[0, 1]$. However, a human user may be uncertain about various parts of the argument. For example, let $A_1 = \langle\{a, a \rightarrow b\}, b\rangle$, and $\tau(A_1) = 0.2$. This value does not indicate whether the user's uncertainty is on the conclusion of the argument (e.g., b) or in parts of its premises (e.g., a or $a \rightarrow b$). Future work will look into a more nuanced argument uncertainty specification.

Another limitation of our current framework is the simplified representation of arguments as logical propositions. Real-world arguments often involve more complex structures and reasoning patterns, such as analogies, causal reasoning, and appeals to emotion. Capturing these rich argumentation dynamics within a computational framework is a significant challenge that requires further research at the intersection of argumentation theory, natural language processing, and knowledge representation.

In conclusion, the framework and study presented in this paper represent an initial step towards the development of adaptive, human-centric argumentation systems. While much work remains to be done to refine and validate our approach, we believe that this research direction has the potential to enhance the effectiveness of human-AI interaction.

Acknowledgments

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

References

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge - A Logical Approach to Probabilities*. MIT Press.

Belesiotis, A.; Rovatsos, M.; and Rahwan, I. 2010. Agreeing on Plans through Iterated Disputes. In *Proceedings of International Conference on Agents and Multiagent Systems (AAMAS)*, 765–772.

Bench-Capon, T. J.; and Dunne, P. E. 2007. Argumentation in Artificial Intelligence. *Artificial Intelligence*, 171(10-15): 619–641.

Besnard, P.; and Hunter, A. 2014. Constructing Argument Graphs with Deductive Arguments: A Tutorial. *Argument & Computation*, 5(1): 5–30.

Budán, M. C.; Cobo, M. L.; Martinez, D. C.; and Simari, G. R. 2020. Proximity Semantics for Topic-based Abstract Argumentation. *Information Sciences*, 508: 135–153.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 156–163.

Collins, A.; Magazzeni, D.; and Parsons, S. 2019. Towards an Argumentation-based Approach to Explainable Planning. In *Proceedings of the International Workshop on Explainable Planning (XAIP)*, 16.

Dennis, L. A.; and Oren, N. 2022. Explaining BDI Agent Behaviour through Dialogue. *Autonomous Agents and Multi-Agent Systems*, 36(2): 29.

Dung, H. T.; and Son, T. C. 2022. On Model Reconciliation: How to Reconcile When Robot Does not Know Human's Model? In *Proceedings of the International Conference on Logic Programming (ICLP)*, 27–48.

Dung, P. M.; and Thang, P. M. 2010. Towards (Probabilistic) Argumentation for Jury-based Dispute Resolution. In *Proceedings of the International Conference on Computational Models of Argument (COMMA)*, 171–182.

Fan, X. 2018. On Generating Explainable Plans with Assumption-Based Argumentation. In *Proceedings of the Principles and Practice of Multi-Agent Systems (PRIMA)*, 344–361.

Fan, X.; and Toni, F. 2015. On Computing Explanations in Argumentation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 1496–1502.

Fox, C. R.; and Poldrack, R. A. 2009. Prospect Theory and the Brain. In *Neuroeconomics*, 145–173.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable Planning. *arXiv preprint arXiv:1709.10256*.

Gonzalez, R.; and Wu, G. 1999. On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38(1): 129–166.

Gordon, T. F. 1994. An Inquiry Dialogue System. *Artificial Intelligence and Law*, 2: 239–292.

Hunter, A. 2013. A Probabilistic Approach to Modelling Uncertain Logical Arguments. *International Journal of Approximate Reasoning*, 54(1): 47–81.

Hunter, A. 2014. Probabilistic Strategies in Dialogical Argumentation. In *Proceedings of the International Conference on Scalable Uncertainty Management (SUM)*, 190–202.

Hunter, A. 2022. Argument Strength in Probabilistic Argumentation based on Defeasible Rules. *International Journal of Approximate Reasoning*, 146: 79–105.

Kahneman, D.; and Tversi, A. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2): 263–292.

Kelley, C. T. 2003. *Solving Nonlinear Equations with Newton's Method*. SIAM.

Li, H.; Oren, N.; and Norman, T. J. 2011. Probabilistic Argumentation Frameworks. In *Proceedings of the International Workshop on Theory and Applications of Formal Argumentation (TAFA)*, 1–16.

Nielsen, S. H.; and Parsons, S. 2006. A Generalization of Dung's Abstract Framework for Argumentation: Arguing with Sets of Attacking Arguments. In *Proceedings of the International Workshop on Argumentation in Multi-Agent Systems (ArgMAS)*, 54–73.

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.

Prakken, H. 2006. Formal Systems for Persuasion Dialogue. *The Knowledge Engineering Review*, 21(2): 163–188.

Rago, A.; Li, H.; and Toni, F. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proceedings of the International Conference on Knowledge Representation and Reasoning (KR)*, 582–592.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Pearson.

Saraswat, D.; Bhattacharya, P.; Verma, A.; Prasad, V. K.; Tanwar, S.; Sharma, G.; Bokoro, P. N.; and Sharma, R. 2022. Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*, 10: 84486–84517.

Setchi, R.; Dehkordi, M. B.; and Khan, J. S. 2020. Explainable Robotics in Human-Robot Interactions. *Procedia Computer Science*, 176: 3057–3066.

Shams, Z.; De Vos, M.; Oren, N.; and Padget, J. 2016. Normative Practical Reasoning via Argumentation and Dialogue. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1244–1250.

Son, T. C.; Nguyen, V.; Vasileiou, S. L.; and Yeoh, W. 2021. Model Reconciliation in Logic Programs. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 393–406.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 518–526.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of Explanations as Model Reconciliation. *Artificial Intelligence*, 301: 103558.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2018. Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4829–4836.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2021. Using State Abstractions to Compute Personalized Contrastive Explanations for AI Agent Behavior. *Artificial Intelligence*, 301: 103570.

Student. 1908. The Probable Error of a Mean. *Biometrika*, 1–25.

Tversky, A.; and Kahneman, D. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4): 297–323.

Vasileiou, S. L.; Kumar, A.; Yeoh, W.; Son, T. C.; and Toni, F. 2023. DR-HAI: Argumentation-based Dialectical Reconciliation in Human-AI Interactions. *arXiv preprint arXiv:2306.14694*.

Vasileiou, S. L.; Previti, A.; and Yeoh, W. 2021. On Exploiting Hitting Sets for Model Reconciliation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 6514–6521.

Vasileiou, S. L.; and Yeoh, W. 2023. PLEASE: Generating Personalized Explanations in Human-Aware Planning. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2411–2418.

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.

Zhong, Q.; Fan, X.; Toni, F.; and Luo, X. 2014. Explaining Best Decisions via Argumentation. In *Proceedings of the European Conference on Social Intelligence (ECSI)*, 224–237.