Does Your AI Agent Get You? A Personalizable Framework for Approximating Human Models from Argumentation-based Dialogue Traces

Yinxu Tang, Stylianos Loukas Vasileiou, William Yeoh

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

Abstract

Explainable AI is increasingly employing argumentation methods to facilitate interactive explanations between AI agents and human users. While existing approaches typically rely on predetermined human user models, there remains a critical gap in dynamically learning and updating these models during interactions. In this paper, we present a framework that enables AI agents to adapt their understanding of human users through argumentation-based dialogues. Our approach, called *Persona*, draws on prospect theory and integrates a probability weighting function with a Bayesian belief update mechanism that refines a probability distribution over possible human models based on exchanged arguments. Through empirical evaluations with human users in an applied argumentation setting, we demonstrate that Persona effectively captures evolving human beliefs, facilitates personalized interactions, and outperforms state-of-the-art methods.

Resources — https://github.com/YODA-Lab/Persona

Introduction

As AI systems become more integrated into real-world applications, the need for transparency and trust in human-AI interactions grows. Explainable AI (XAI) addresses this need by focusing on generating understandable explanations for human users that foster trust and accountability (Gunning and Aha 2019). A key paradigm within XAI is argumentation (Čyras et al. 2021), which enables interactive, dialogue-based explanation processes between AI agents and human users. These processes offer improved clarity and foster stronger human-AI interactions.

A core assumption in most existing argumentation-based XAI work is that the AI agent has a static, deterministic model of the human user that it uses in its deliberative processes. While assuming the AI agent has access to an a-priori human model has its advantages (Sreedharan, Chakraborti, and Kambhampati 2021; Vasileiou, Previti, and Yeoh 2021), this approach often falls short of capturing the intricate complexities of real-world interactions. It is not only likely that humans hold beliefs at different levels of granularity and with varying degrees of certainty, but also that their beliefs evolve dynamically over time. Such simplifications can lead to significant misalignments between AI agents and human users (Russell 2019), as the AI agent might base its decisions or explanations on an inaccurate or incomplete understanding of the human.

As a step towards addressing this issue, in this paper, we propose a novel approach that enables AI agents to adapt their decisions and explanations based on a dynamic understanding of human mental states. Our method represents human models as probability distributions that are continuously refined through ongoing argumentative interactions, building upon established frameworks in computational argumentation (Gordon 1994; Parsons, Wooldridge, and Amgoud 2003; Prakken 2006; Hunter 2015, 2016; Rago, Li, and Toni 2023; Vasileiou et al. 2024). Our proposed framework, *Personalized Human Model Approximations* (Persona), integrates two key components to achieve this goal. First, it employs a *Bayesian belief update* mechanism that systematically refines the human model based on observed interaction patterns. Second, it incorporates a probability weighting function derived from *prospect theory* (Tversky and Kahneman 1992), which accounts for human tendencies to overweight low probabilities and underweight high probabilities in decision-making contexts. This dual approach enables Persona to offer personalized interactions by capturing individual differences in how users evaluate probabilistic information during argumentative exchanges.

Furthermore, we conduct an extensive evaluation of our approach using real argumentation-based dialogue traces collected via a human-subject study. Our results show that Persona not only effectively captures and updates human models, but it also outperforms existing state-of-the-art argumentation-based methods.

The main contributions of this paper are as follows:

- We introduce *Persona*, a novel framework for approximating and updating a probabilistic human model through argumentation-based dialogue traces. Our framework incorporates a prospect-theory-inspired probability weighting function with a Bayesian belief update mechanism.
- We conduct a human-subject study on an argumentation-based dialogue scenario and collect dialogue traces involving human users. We empirically evaluate the effectiveness of our approach on these traces and demonstrate its ability to capture evolving human models and facilitate

personalized interactions, while also outperforming state-of-the-art methods.

Related Work

Argumentation-based Dialogues

According to the influential work by 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. Related dialogue types include: Persuasion (Gordon 1994; Prakken 2006), where an agent attempts to convince another agent to accept a proposition they initially do not hold; information-seeking (Parsons, Wooldridge, and Amgoud 2003; Fan and Toni 2012), where an agent seeks to obtain information from another agent believed to possess it; and inquiry (Hitchcock and Hitchcock 2017; Black and Hunter 2009), where two agents collaborate to find a joint proof for a query that neither could prove individually. The advent of argumentative dialogue-based systems (Black, Maudet, and Parsons 2021) illustrates the great potential of argumentation for collaborative decision-making and consensus-building in human-AI interaction settings. However, these approaches often neglect the dynamic nature of belief updating during dialogues.

On a similar thread, our work fits well within the literature on argumentation-based explainable AI (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; Vasileiou et al. 2024). While these approaches provide a solid foundation for argumentation-based explanations, they do not explicitly focus on approximating the human users model, which is central to this paper.

Human Model Approximation

Accurate human models are crucial for effective human-AI interactions. In argumentation, several approaches have emerged. Rienstra, Thimm, and Oren (2013) proposed a probabilistic opponent model for move selection based on perceived awareness. Hadjinikolis et al. (2013) explored dialogue history analysis to predict opponent arguments. Hadoux et al. (2015) introduced probabilistic finite state machines and partially observable Markov decision processes for modeling dialogue progression under uncertainty.

In other domains, various approaches to human model approximation exist. Deep learning has been used to simulate and predict human behavior from large datasets (Hamrick 2019; Lake et al. 2017). Game-theoretic models reveal how agents' mental states affect choices and strategies in competitive scenarios (Yoshida, Dolan, and Friston 2008; Camerer 2011). Planning formalisms have been utilized to learn human models in human-AI interaction settings (Sreedharan, Chakraborti, and Kambhampati 2018; Sreedharan, Srivastava, and Kambhampati 2018; Black, Coles, and Bernardini 2014). The problem of learning human preferences has also been extensively studied, particularly in recommendation systems (Fürnkranz and Hüllermeier 2010). Preferences are often elicited via ranking or comparisons (Ailon 2012; Wirth et al. 2017), or reinforcement learning paradigms (Wilson, Fern, and Tadepalli 2012; Bıyık, Talati, and Sadigh 2022).

Most relevant to our work are those by Hunter (2013, 2015, 2016), which present methods for representing and updating human beliefs through probability distributions during persuasion dialogues. While they provided essential theoretical groundwork, our approach extend them by incorporating insights from prospect theory and introducing personalized modeling capabilities that account for individual differences in probability assessment.

As our approach is specifically designed for approximating human models in argumentation-based dialogues, we compare it to the most relevant work in this space, namely the work by Hunter (2015, 2016). We do not compare it against non-argumentation approaches, as they lack the specific structures and mechanisms necessary for handling structured arguments and belief updates in dialogue settings. Our focus on argumentative reasoning and uncertainty in dialogues requires specialized techniques that these general approaches do not provide.

Background

We will use classical propositional logic to describe aspects of the world. 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 all models of ϕ.

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

Definition 1 (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, the probability distribution allows an agent to create a *ranking* between possible words with respect to how likely they are to be true. This then allows the agent to compute the *probability of a formula* as follows:

Definition 2 (Probability of Formula)**.** *Let $\mathcal{M}$ be the set of models of language $\mathcal{L}$ and P a probability distribution over $\mathcal{M}$. The probability of formula $\phi \in \mathcal{L}$ is $P(\phi) = \sum_{m \models \phi} P(m)$.*

Argumentation-based Dialogues: In an argumentation-based dialogue, agents take turns exchanging arguments that prove (or disprove) specific claims, where the structure and relationships between these arguments are governed by the underlying argumentation semantics (Black, Maudet, and Parsons 2021). In this paper, we consider the semantics of structured (deductive) argumentation (Besnard and Hunter 2014), where each argument is constructed using formulae from language $\mathcal{L}$. Formally,

Definition 3 (Argument)**.** *Let $\mathcal{L}$ be the language and $\phi \in \mathcal{L}$ a formula. Then,*

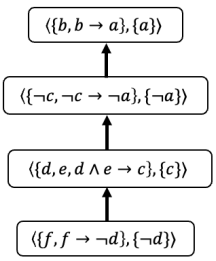

Figure 1: Example dialogue trace $\mathcal{T}$ between two agents.

- ***Argument:*** *$A = \langle \Phi, \phi \rangle$ is an argument for ϕ iff: (i) $\Phi \subseteq \mathcal{L}$; (ii) $\Phi \models \phi$; (iii) $\Phi \not\models \bot$; and (iv) $\nexists \Phi' \subset \Phi$ s.t. $\Phi' \models \phi$. We refer to ϕ as the* claim *of the argument, and Φ as the* premise *of the argument.*
- ***Attack Relation:*** *An argument $A' = \langle \Psi, \psi \rangle$* attacks *argument $A = \langle \Phi, \phi \rangle$ and vice versa (i.e., A attacks A') iff $\Phi \cup \Psi \models \bot$, where $\bot$ denotes falsity.*

The attack relation captures conflicts between arguments, which is essential for modeling disagreements.

Example 1. *Let $\mathcal{L}$ be the language consisting of variables $\{a, b, c\}$. Then, $A_1 = \langle \{b, b \rightarrow a\}, a \rangle$ and $A_2 = \langle \{\neg c, \neg c \rightarrow \neg a\}, \neg a \rangle$ are two arguments for a and $\neg a$, respectively. Note that A_1 attacks A_2 and vice versa, as $\{b, b \rightarrow a, \neg c, \neg c \rightarrow \neg a\} \models \bot$.*

In real-world scenarios, arguments often come with some degree of uncertainty. We can capture this uncertainty with a probability distribution over the models $\mathcal{M}$ of $\mathcal{L}$, and then use it to compute the probability for any argument $A = \langle \Phi, \phi \rangle$ using Definition 2, i.e., $P(A) = \sum_{m \models A} P(m)$, where $m \models A$ is a shorthand notation to mean that the premise Φ of A is true in m.

Now, in this paper, we are mainly interested in argumentation-based *dialogue traces* between two agents, i.e., finite sequences of arguments that attack each other:

Definition 4 (Dialogue Trace)**.** *Let Δ be an argumentation-based dialogue between agents α and η. A dialogue trace from Δ is defined as $\mathcal{T} = \langle (A_1, x_1)^{t_1}, (A_2, x_2)^{t_2}, \ldots, (A_n, x_n)^{t_n} \rangle$, where each $(A_i, x_i)^{t_i}$ denotes the argument put forward by agent $x_i \in \{\alpha, \eta\}$ at timestep t_i.*

A dialogue trace $\mathcal{T}$ can also be represented as a tree with n nodes and $n - 1$ edges, where each node i represents the argument expressed at timestep t_i, and there is a directed edge from node j to node i iff argument A_j attacks argument A_i, where $1 \leq i < j \leq n$. Note that repeating arguments within the same dialogue trace are not allowed to avoid infinite loops. Figure 1 shows an example of the dialogue trace.

Approximating Human Models

We now introduce our framework that enables an agent to progressively update its approximation of the human model through argumentation-based dialogue traces.

Problem Setting and Assumptions: Our setting consists of an AI agent (denoted α) interacting with a human user (denoted η) via an argumentation-based dialogue. We make the following key assumptions:

- **Shared Domain Language**: Both α and η have access to and communicate in the same language $\mathcal{L}$ using a shared vocabulary of atomic variables. This allows them to construct domain-specific formulae.
- **Probabilistic Human Model**: The human model is represented as a probability distribution $P_h^{t_i}$ over the possible models $\mathcal{M}$ of $\mathcal{L}$ at each timestep t_i. Initially, we assume a uniform distribution: $P_h^{t_0}(m) = \frac{1}{|\mathcal{M}|}$ for all $m \in \mathcal{M}$.
- **Dialogue Traces**: We have access to (finite) dialogue traces $\mathcal{T}$ produced by argumentation-based dialogues between α and η (Vasileiou et al. 2024).

In real-world argumentation, arguments often come with some degree of uncertainty. To capture this, we associate a probability $p(A_i)$ with each argument A_i in the dialogue trace. It is crucial to note that these probabilities represent uncertainty from the perspective of the human user, that is, how likely the human thinks that the argument is true.[1]

Updating the Human Model

Given a dialogue trace $\mathcal{T}$, we employ a Bayesian approach to update the agent's probability distribution P_h over possible human models. At each timestep t_i, when an argument A_i is presented, we perform the following update:

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

This update mechanism 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.

A More Personalized Approach to Uncertainty Estimation: While the Bayesian update approach provides a solid foundation for estimating the human model, it does not account for the subjective nature of how humans perceive and think about uncertainty. To address this, we introduce a more personalized approach based on prospect theory (Kahneman and Tversi 1979), which allows us to capture individual differences in how humans evaluate probabilities in argumentative contexts.[2]

We propose the following probability weighting function to model the relationship between "actual" probabilities

[1] These probabilities can arise from various sources, such as incomplete or imprecise knowledge, subjective interpretations, or lack of confidence in the reasoning process. Accounting for these uncertainties is crucial for developing a more realistic and nuanced model of human reasoning in argumentative contexts.

[2] According to prospect theory, humans tend to overweight small probabilities and underweight large probabilities.

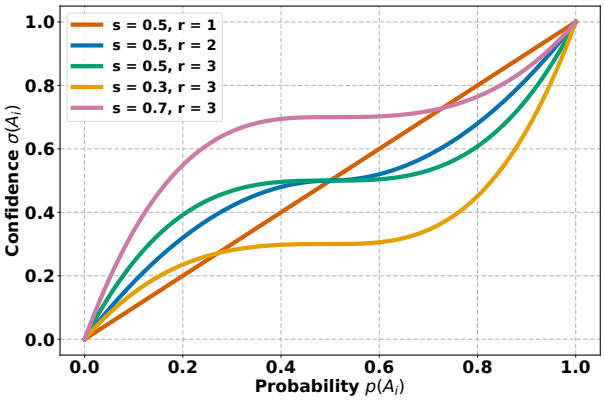

Figure 2: Probability weighting function with value pairs (s, r) given as $\{(0.5, 1), (0.5, 2), (0.5, 3), (0.3, 3), (0.7, 3)\}$. Lower values of s imply that the average probability reflects a lower level of human confidence in the agent's argument, whereas higher values of r suggest excessive distortion, either through overweighting or underweighting of the probability. Note that the relationship between confidence and probability is linear when $s = 0.5$ and $r = 1$.

$p(A_i)$ and subjective probability perceptions, referred to as *confidence values* $\sigma(A_i)$:

$$\sigma(A_i) = \begin{cases} s + (1-s) \cdot (2 \cdot p(A_i) - 1)^r & \text{if } p(A_i) > 0.5 \\ s - s \cdot (1 - 2 \cdot p(A_i))^r & \text{if } p(A_i) \leq 0.5 \end{cases} \tag{2}$$

where parameter $s \in (0, 1)$ determines the value of $\sigma(A_i)$ when $p(A_i) = 0.5$, regardless of the value of r, and parameter $r \in [1, \infty)$ controls the degree of this nonlinear distortion. The relationship between s and r is shown in Figure 2.

In practice, we often need to infer the probability $p(A_i)$ from the observed subjective probability $\sigma(A_i)$. To do this, we invert Equation 2:

$$p(A_i) = \begin{cases} \frac{1}{2} - \frac{1}{2} \cdot (\frac{s - \sigma(A_i)}{s})^{\frac{1}{r}} & \text{if } \sigma(A_i) \leq s \\ \frac{1}{2} + \frac{1}{2} \cdot (\frac{\sigma(A_i) - s}{1-s})^{\frac{1}{r}} & \text{if } \sigma(A_i) > s \end{cases} \tag{3}$$

Example 2. *Consider the dialogue trace shown in Figure 1. At timestep t_1, the agent asserts the argument $A_1 = \langle\{b, b \to a\}, \{a\}\rangle$. The human assigns a confidence value of $\sigma(A_1) = 0.6$ to this argument. Assuming $s = 0.5$ and $r = 1.5$, the actual probability of A_1 is computed using Equation 3:*

$$p(A_1) = \frac{1}{2} + \frac{1}{2} \cdot (\frac{0.6 - 0.5}{1 - 0.5})^{\frac{1}{1.5}} \approx 0.67$$

Suppose there are eight possible models, $\mathcal{M} = \{m_1, m_2, \ldots, m_8\}$, with a uniform prior distribution $P_h^{t_0}(m_1) = \ldots = P_h^{t_0}(m_8) = 0.125$. Let m_1 and m_2 be the models that entail argument A_1, i.e., $m_1, m_2 \models A_1$. Applying the update mechanism from Equation 1, we get:

$$P_h^{t_1}(m_1) = P_h^{t_1}(m_2) = \frac{0.125}{0.125 + 0.125} \cdot 0.67 = 0.335$$
$$P_h^{t_1}(m_3) = P_h^{t_1}(m_4) = P_h^{t_1}(m_5) = P_h^{t_1}(m_6) = P_h^{t_1}(m_7)$$
$$= P_h^{t_1}(m_8) = \frac{0.125}{0.125 \cdot 6} \cdot 0.33 = 0.055$$

After this update, the models m_1 and m_2 that are consistent with the agent's argument have a higher probability than the other six models, reflecting the human's moderate confidence in the argument.

At the next timestep t_2, the human presents the argument $A_2 = \langle\{\neg c, \neg c \to \neg a\}, \{\neg a\}\rangle$ with probability $p(A_2) = 0.9$. Let m_3 and m_4 be the models that entail argument A_2. Applying the update mechanism again, we get:

$$P_h^{t_2}(m_1) = P_h^{t_2}(m_2) = \frac{0.335}{0.335 \cdot 2 + 0.055 \cdot 4} \cdot 0.1 = 0.038$$
$$P_h^{t_2}(m_3) = P_h^{t_2}(m_4) = \frac{0.055}{0.055 + 0.055} \cdot 0.9 = 0.45$$
$$P_h^{t_2}(m_5) = P_h^{t_2}(m_6) = P_h^{t_2}(m_7) = P_h^{t_2}(m_8)$$
$$= \frac{0.055}{0.335 \cdot 2 + 0.055 \cdot 4} \cdot 0.1 = 0.006$$

After this update, the models m_3 and m_4 that are consistent with the human's argument have a much higher probability than the models consistent with the agent's previous argument. The same process can be applied in the remaining two timesteps.

Personalized Parameter Learning: To adapt the personalization parameters s and r to individual users, we can use a data-driven approach based on dialogue traces and user-provided model rankings. This method aims to find the optimal parameters that maximize the correlation between our computed model rankings and the ground truth rankings provided by users.

The approach involves collecting two types of data from users: Dialogue traces and model rankings. For a given pair of parameters (s, r), we compute a ranking over models using the probability weighting function and belief update mechanism described earlier. We then evaluate the fit of parameters (s, r) by computing the correlation between our computed ranking and the user-provided ground truth ranking. The optimal parameters (s^*, r^*) are then determined by maximizing the correlation:

$$(s^*, r^*) = \underset{(s,r)}{\text{argmax}}\, \rho(s, r) \tag{4}$$

where $\rho(s, r)$ denotes the correlation between computed and ground truth rankings.

The specific implementation details, including how we split the data for learning and evaluation, is discussed in the empirical evaluation section.

Empirical Evaluations

We now evaluate the effectiveness of our approach across two dimensions: (1) Its ability to personalize and optimize the probability weighting function parameters; and (2) Its performance in approximating human models and estimating argument probabilities compared to existing methods. To collect data, we conducted the following human-subject study.

Human-Subject Study Description

We simulated a scenario where participants interacted with an AI assistant named *Blitzcrank* to evaluate the suitability of a fictional venue, *Luminara Gardens*, for a company team-building event. This scenario was chosen to provide a concrete context for argumentation while being accessible to a general participant pool.

The study consisted of a series of interaction rounds (maximum 5) between each participant and Blitzcrank. Each round followed this structure:

- Blitzcrank presented an argument about Luminara Gardens' suitability.
- Participants rated their confidence in Blitzcrank's argument on a five-point scale: Very low (0.1), low (0.3), average (0.5), high (0.7), or very high (0.9).
- Participants selected and presented a counterargument to Blitzcrank from a set of three options, each associated with a confidence level.
- Participants ranked four different perspectives (i.e., models) on Luminara Gardens' suitability.

The dialogue continued for up to five rounds, with the option to end earlier if agreement was reached.

Data Collection: We recruited 200 participants via the Prolific platform (Palan and Schitter 2018), ensuring a diverse sample.[3] Participants were required to be fluent in English and were compensated USD 4.00 for their time. After applying attention checks and coherence filters, we retained data from 184 participants for analysis. For each participant i, we collected:

- Dialogue traces $\mathcal{T}_i = \langle (A_1, x_1, \sigma_1)^{t_1}, \ldots, (A_{n_i}, x_{n_i}, \sigma_{n_i})^{t_{n_i}} \rangle$, where $n_i \in \{8, 10\}$ is the number of completed interactions, $x_j \in \{\text{Blitzcrank}, \text{Participant}\}$, and σ_j is the participant's confidence value on argument A_j.
- Model rankings $M_i^t = \langle m_1^t, m_2^t, m_3^t, m_4^t \rangle$ after each round t, where each round consists of two interactions (e.g., two exchanged arguments).
- Final argument rankings $R_i = \langle a_1, a_2, \ldots, a_m \rangle$, where m is the total number of arguments presented.
- Post-study questionnaire responses assessing satisfaction and interaction quality.

Experiment 1: Learning Optimal Personalization Parameters

Our first experiment aimed to learn the optimal values for the personalization parameters s and r in our probability weighting function (Equations 2 and 3). This data-driven approach uses dialogue traces and user-provided model rankings to maximize the correlation between our computed model rankings and the ground truth rankings provided by the participants.

Methodology: For each participant i with n_i interactions, we performed the following steps: First, we iterated over each $k \in \{1, 2, 3\}$ rounds of interactions. Then, for each k, we learn the optimal (s_i^{k*}, r_i^{k*}) for participant i by:

$$(s_i^{k*}, r_i^{k*}) = \underset{(s', r')}{\text{argmax}} \sum_{t=1}^{k} \rho(M_i^t, \hat{M}_i^t(s', r')) \quad (5)$$

where ρ is Spearman's rank correlation coefficient (Spearman 1904), M_i^t is the participant's model ranking at t, and $\hat{M}_i^t(s', r')$ is the computed ranking using parameters (s', r'), where $s' \in \{0.1, 0.2, \ldots, 0.9\}$ and $r' \in \{1, 2, \ldots, 8\}$. Specifically, to compute $\hat{M}_i^t(s', r')$, we:

- Used Equation 3 to transform confidence values σ_j in the dialogue trace to probabilities.
- Applied the belief update mechanism (Equation 1) to compute the distribution $P_h^t(m)$ over models.
- Ranked the models based on their probabilities in $P_h^t(m)$.

We then evaluated the learned optimal values (s_i^{k*}, r_i^{k*}) for each participant i in a future round $k' > k$:

$$\rho_i^{k'} = \rho(M_i^{k'}, \hat{M}_i^{k'}(s_i^{k*}, r_i^{k*})) \quad (6)$$

This approach allows us to assess how well the learned parameters generalize to new, unseen interactions in round k'. By varying k, we can analyze how the amount of training data affects the model's performance.

Evaluation Metrics: We use Spearman's Rank Correlation ρ between computed and ground truth rankings, and Paired Student's t-tests to assess the statistical significance between the evaluated methods.

Results: Figure 3 shows the distribution of Spearman's rank correlation coefficients in round four (i.e., $k' = 4$ in Equation 6) using the optimized parameters from the first k interaction data $\mathcal{D}_k$. We applied the learned parameters in round four since the minimum number of rounds for all participants was four. A high positive correlation indicates method effectiveness, with particular attention given to distributions

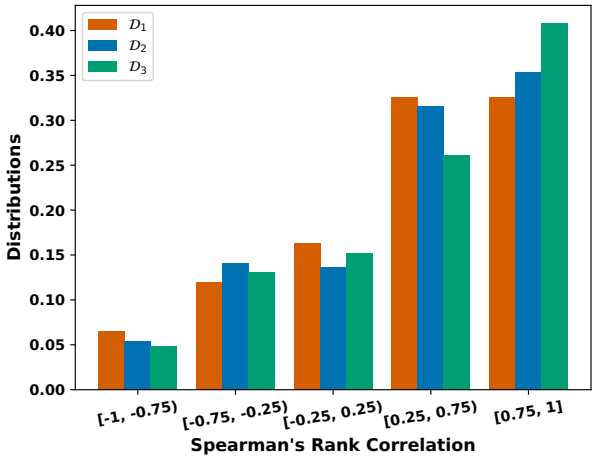

Figure 3: Comparisons of Spearman's rank correlation distributions in round four derived from the first k rounds of interaction data $\mathcal{D}_k$.

[3]Ethics approval was obtained through our university's IRB.

X \ Y	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
$\mathcal{D}_1$	–	0.925	1
$\mathcal{D}_2$	0.075	–	0.985
$\mathcal{D}_3$	4.43×10^{-4}	0.015	–

Table 1: The p-values from Student's t-tests assessing the hypothesis that X outperforms Y in Experiment 1.

above 0.75. As the value of k increases, our approach can better approximate the human model for the fourth round by leveraging more data from previous rounds, enhancing parameter personalization for each participant. In this way, the more data we use to learn personalized parameters, the more accurately the human model is approximated.

To further assess the personalization component of Persona with different values of k, we conducted paired Student's t-tests. Table 1 presents the p-values, evaluating the hypothesis that X (rows of the table) outperforms Y (columns). The results show that $\mathcal{D}_3$ statistically significantly outperforms $\mathcal{D}_1$ and $\mathcal{D}_2$, with p-values smaller than 0.05. While $\mathcal{D}_2$ does outperform $\mathcal{D}_1$, there is no statistical significance, as the p-value is greater than 0.05.

Experiment 2: Comparative Evaluation

The goal of our second experiment is two-fold: (1) To evaluate the effectiveness of our approach on approximating human models; and (2) To evaluate the effectiveness of our approach on estimating the human beliefs of arguments. We used the same evaluation metrics as in Experiment 1.

Experiment 2.1: Human Model Approximation

In this experiment, we evaluated the efficacy of our personalized approach, referred to as *Persona* in subsequent figures, in approximating human models. We compared our method against the following baselines:[4]

- *Generic*: Instead of personalizing parameters for each participant, we learned the same (s, r) for each participant in the first k rounds, i.e., Equation 5 can be modified as:

$$(s^{k*}, r^{k*}) = \underset{(s',r')}{\operatorname{argmax}} \sum_i \sum_{t=1}^{k} \rho(M_i^t, \hat{M}_i^t(s', r')) \quad (7)$$

  This serves as an ablation study for Persona.

- *SBU*: The simple Bayesian update we proposed in Equation 1. This serves as an ablation study for Persona as well.

- HM_1: An argumentation-based method for updating probability distributions of human models based on argument graphs (Hunter 2015).

- HM_2: An enhanced version of Hunter's HM_1 that utilizes the argument structure for updating the distribution (Hunter 2015).

[4]Please refer to the supplement in our GitHub repository for details about the baselines.

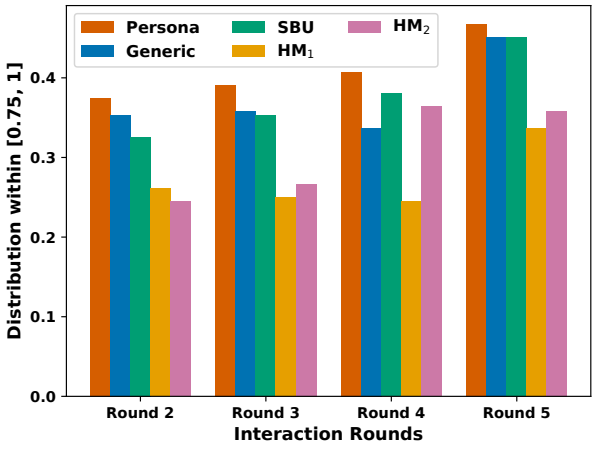

Figure 4: Comparisons of Spearman's rank correlation distributions in model estimation within $[0.75, 1]$ in round k ($k = 2, 3, 4, 5$) of human model rankings where parameters are learned from the previous $k-1$ rounds. Note that for participants with only four interactions, the results for Round 5 are identical to those of Round 4.

X \ Rounds	Round 2	Round 3	Round 4	Round 5
Generic	0.043	0.004	0.010	0.640
SBU	0.047	0.003	0.047	0.426
HM_1	2.408×10^{-6}	3.255×10^{-4}	3.760×10^{-5}	0.006
HM_2	5.730×10^{-5}	0.002	0.006	0.001

Table 2: The p-values from Student's t-tests assessing the hypothesis that Persona outperforms X in Experiment 2.1.

Results: We compared the Spearman's rank correlation distributions in round $k = \{2, 3, 4, 5\}$ of human model rankings where parameters are learned from the previous $k-1$ rounds among Persona and its ablations and the two baselines. Figure 4 displays the distribution of Spearman's rank correlation coefficients for interval $[0.75, 1]$ for human models.[5] We observed that Persona performed better than all the other methods in all rounds. Compared to *Generic* and *SBU*, the results demonstrate that incorporating both personalization and the weighting function increases the accuracy of model approximation. Notably, Persona significantly outperformed HM_1 and across all rounds, and HM_2 across rounds 2, 3, and 5. Interestingly, Hunter's HM_2 has close results to Persona in round 4 due to the randomness of the method during the ranking procedure.

We also conducted paired Student's t-tests among various methods, where Table 2 presents the p-values evaluating the hypothesis that Persona outperforms method X in human model approximation across different rounds. The results demonstrate that Persona statistically significantly outperforms all the other methods in almost all rounds. These findings underscore Persona's capacity to effectively utilize existing data to learn personalized parameters for each participant, thereby enhancing the accuracy of human model

[5]We omit Spearman's rank correlation coefficients in other intervals, but we describe them in the supplement.

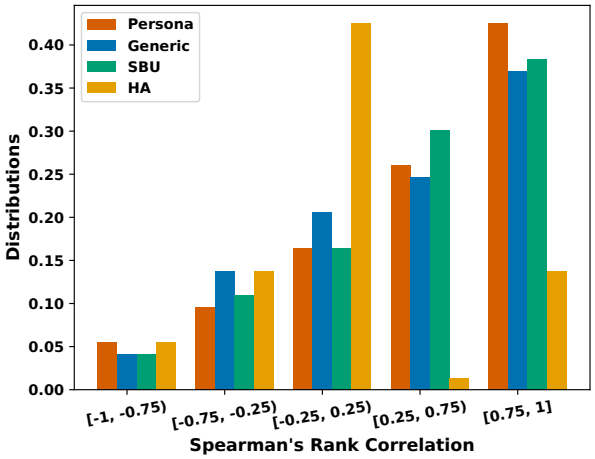

Figure 5: Comparisons of Spearman's rank correlation distributions for argument beliefs across different methods.

estimation in the subsequent round compared to state-of-the-art baselines and the ablation variants.

Experiment 2.2: Human Argument Belief Estimation
In this experiment, we evaluated the efficacy of Persona in estimating argument beliefs. For baselines, we used the state-of-the-art method for learning probability distributions of arguments by Hunter (2016), referred to as HA, as well as the ablation variants used in Experiment 2.1. Unlike in Experiment 2.1, we only use data from a subset of individuals from the human-subject study. Specifically, we omitted data of individuals whose dialogue traces ended with the agent because, for these individuals, there were only *two* relevant arguments that they needed to rank. In contrast, the other group of individuals who ended their conversations had *four* arguments that they needed to rank. Two arguments are, in our opinion, too few for meaningful comparisons.

Results: Figure 5 displays the distribution of Spearman's rank correlation coefficients for argument beliefs. Persona applies the personalized values of s and r during the first $n_i - 1$ rounds for each participant i in argument belief estimation, while *Generic* uses fixed values of s and r learned from the first three rounds across all participants. Focusing on the range $[0.75, 1]$ of high correlation, our results show that Persona outperforms its ablation variants, demonstrating the benefits of learning and personalization in improving argument belief estimations. Additionally, it surpasses the HA method as well.

To better understand the statistical significance of this observation, we also conducted paired Student's t-tests. Table 3 presents the p-values evaluating the hypothesis that method X (rows) outperforms method Y (columns) in argument belief approximation. Surprisingly, there is no statistical difference between Persona and *SBU*, as the p-value is significantly larger than 0.05. The reason is that while Persona does better than *SBU* in the high correlation range, *SBU* does better in the other ranges. However, the improvement of Persona over *Generic* and HA are statistically significant, with p-values smaller than 0.05. Additionally, especially noteworthy are the extremely small p-values for Persona and the two ablation variants over HA, highlighting the strength of our framework against the state of the art.

X \ Y	*Persona*	*Generic*	*SBU*	*HA*
Persona	–	0.002	0.569	2.105×10^{-11}
Generic	0.998	–	0.999	5.674×10^{-8}
SBU	0.431	0.001	–	1.410×10^{-11}
HA	1	1	1	–

Table 3: The p-values from Student's t-tests assessing the hypothesis that X outperforms Y in Experiment 2.2.

Computational Results

We implemented Persona and evaluated its performance on a MacBook Pro with a 2.2 GHz Quad-Core Intel Core i7 processor and 16GB of RAM. Persona took approximately 0.6 seconds to compute probabilities for each s and r pair of hyperparameter values per participant. In comparison, methods HM_1 and HM_2 took around 9 seconds and 41 seconds per participant, respectively, whereas HA required just 0.003 seconds to compute argument beliefs. These runtimes indicate that all approaches, particularly Persona, are suitable for real-time evaluations in practical applications. However, it is important to note that additional time would be required for translating between natural language and logic, which is an area we plan to address in future work.

Conclusions and Future Work

In this paper, we introduced Persona, a novel framework for personalizing human model approximations in argumentation-based dialogues. Persona combines a Bayesian belief update mechanism that refines probability distributions over potential human models during dialogues with a prospect theory-inspired probability weighting function. This combination allows for the incorporation of uncertainty estimates for both agent and human arguments while capturing individual differences in how humans evaluate probabilities in argumentative contexts.

Through a comprehensive human-subject study involving 184 participants, we demonstrated the effectiveness of Persona in both model approximation and argument belief estimation. Our empirical evaluations showed that Persona significantly outperforms state-of-the-art methods in terms of Spearman's rank correlation and statistical significance tests. Furthermore, our computational results indicate that Persona is suitable for practical applications, with competitive runtime performance compared to existing methods.

For future work, we plan to investigate how these learned human models can be used to generate more persuasive arguments as well as apply them to other applications, including automated planning (Chakraborti et al. 2017; Sreedharan, Chakraborti, and Kambhampati 2020; Vasileiou et al. 2022; Vasileiou and Yeoh 2023) and scheduling (Čyras et al. 2019; Agrawal, Yelamanchili, and Chien 2020; Pozanco et al. 2022; Vasileiou, Xu, and Yeoh 2023).

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

Agrawal, J.; Yelamanchili, A.; and Chien, S. 2020. Using Explainable Scheduling for the Mars 2020 Rover Mission. *arXiv preprint arXiv:2011.08733*.

Ailon, N. 2012. An Active Learning Algorithm for Ranking from Pairwise Preferences with an Almost Optimal Query Complexity. *Journal of Machine Learning Research*, 13(1): 137–164.

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

Bıyık, E.; Talati, A.; and Sadigh, D. 2022. Aprel: A Library for Active Preference-based Reward Learning Algorithms. In *Proceedings of the International Conference on Human-Robot Interaction (HRI)*, 613–617.

Black, E.; Coles, A.; and Bernardini, S. 2014. Automated Planning of Simple Persuasion Dialogues. In *Proceedings of the International Workshop on Computational Logic in Multi-Agent Systems (CLIMA)*, 87–104.

Black, E.; and Hunter, A. 2009. An Inquiry Dialogue System. *Autonomous Agents and Multi-Agent Systems*, 19: 173–209.

Black, E.; Maudet, N.; and Parsons, S. 2021. Argumentation-based Dialogue. *Handbook of Formal Argumentation*, 2.

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.

Camerer, C. F. 2011. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

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.

Čyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019. Argumentation for Explainable Scheduling. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2752–2759.

Čyras, K.; Rago, A.; Albini, E.; Baroni, P.; Toni, F.; et al. 2021. Argumentative XAI: A Survey. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4392–4399.

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

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. 2012. Agent Strategies for ABA-based Information-seeking and Inquiry Dialogues. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 324–329.

Fan, X.; and Toni, F. 2015. On Computing Explanations in Argumentation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 344–361.

Fürnkranz, J.; and Hüllermeier, E. 2010. *Preference Learning*. Springer Science & Business Media.

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

Gunning, D.; and Aha, D. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2): 44–58.

Hadjinikolis, C.; Siantos, Y.; Modgil, S.; Black, E.; and McBurney, P. 2013. Opponent Modelling in Persuasion Dialogues. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 164–170.

Hadoux, E.; Beynier, A.; Maudet, N.; Weng, P.; and Hunter, A. 2015. Optimization of Probabilistic Argumentation with Markov Decision Models. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2004–2010.

Hamrick, J. B. 2019. Analogues of Mental Simulation and Imagination in Deep Learning. *Current Opinion in Behavioral Sciences*, 29: 8–16.

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.

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

Hunter, A. 2015. Modelling the Persuadee in Asymmetric Argumentation Dialogues for Persuasion. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 3055–3061.

Hunter, A. 2016. Persuasion Dialogues via Restricted Interfaces using Probabilistic Argumentation. In *Proceedings of the Scalable Uncertainty Management (SUM)*, 184–198.

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

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building Machines that Learn and Think Like People. *Behavioral and Brain Sciences*, 40: e253.

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.

Pozanco, A.; Mosca, F.; Zehtabi, P.; Magazzeni, D.; and Kraus, S. 2022. Explaining Preference-driven Schedules: The EXPRES Framework. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 710–718.

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.

Rienstra, T.; Thimm, M.; and Oren, N. 2013. Opponent Models with Uncertainty for Strategic Argumentation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 332–338.

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

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.

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. 2020. The Emerging Landscape of Explainable Automated Planning & Decision Making. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4803–4811.

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.

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. 2024. Dialectical Reconciliation via Structured Argumentative Dialogues. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 777–787.

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.; Xu, B.; and Yeoh, W. 2023. A Logic-based Framework for Explainable Agent Scheduling Problems. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2402–2410.

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.

Walton, D.; and Krabbe, E. C. 1995. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press.

Wilson, A.; Fern, A.; and Tadepalli, P. 2012. A Bayesian Approach for Policy Learning from Trajectory Preference Queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1133–1141.

Wirth, C.; Akrour, R.; Neumann, G.; and Fürnkranz, J. 2017. A survey of Preference-based Reinforcement Learning Methods. *Journal of Machine Learning Research*, 18(136): 1–46.

Yoshida, W.; Dolan, R. J.; and Friston, K. J. 2008. Game Theory of Mind. *PLoS Computational Biology*, 4(12): e1000254.