Reflective Verbal Reward Design for Pluralistic Alignment*

Carter Blair , Kate Larson , Edith Law
David R. Cheriton School of Computer Science
cblair@uwaterloo.ca

Abstract

AI agents are commonly aligned with "human values" through reinforcement learning from human feedback (RLHF), where a single reward model is learned from aggregated human feedback and used to align an agent's behavior. However, human values are not homogeneous–different people hold distinct and sometimes conflicting values. Aggregating feedback into a single reward model risks disproportionately suppressing minority preferences and unique perspectives. To address this, we present a novel reward modeling approach for learning *individualized* reward models. Our approach uses a language model to guide users through *reflective* dialogues where they critique agent behavior and construct their preferences. This personalized dialogue history, containing the user's reflections and critiqued examples, is then used as context for another language model that serves as an individualized reward function for evaluating new trajectories. In studies with 30 participants, our method achieved a 9-12% improvement in accuracy over non-reflective language-based reward models while being vastly more sample efficient than traditional supervised learning methods.

1 Introduction

As AI systems increasingly permeate everyday life, the need to align them with "human values" has gained urgency. Yet, human values are both varied and context dependent [Schwartz, 1992; Friedman *et al.*, 2013; Le Dantec *et al.*, 2009]. Recent advances in Reinforcement Learning from Human Feedback (RLHF) [Casper *et al.*, 2023; Bai *et al.*, 2022] have relied mainly on training a *single* reward model to capture preferences aggregated from multiple users. By default, these aggregated models implicitly make trade-offs among competing values and can disproportionately suppress minority viewpoints [Siththaranjan *et al.*, 2023; Chakraborty *et al.*, 2024].

In response to this shortcoming, recent work has explored approaches that better preserve the diversity of human values [Siththaranjan *et al.*, 2023; Chakraborty *et al.*, 2024; Poddar *et al.*, 2024]. While the specifics differ, these approaches broadly increase the granularity of the reward signal to account for differences in opinion. For example, some methods learn multiple distinct reward models to capture different preference clusters [Chakraborty *et al.*, 2024]. Others learn reward models that output distributions rather than scalar values to represent uncertainty and variation in preferences [Siththaranjan *et al.*, 2023], and others learn *individualized* reward models for each person [Poddar *et al.*, 2024].

These higher-granularity approaches offer benefits for collective and individual alignment. In collective settings where we must aggregate preferences across many stakeholders, having granular representations of preferences enables making transparent choices about balancing competing values by explicitly aggregating rewards. Rather than implicitly averaging preferences, we can optimize for desired properties like egalitarian welfare. Meanwhile, for personal AI assistants, techniques that learn high granularity preference representations can give rewards that are more representative of the given user compared to the population average. However, a key challenge exists: While a single reward model can leverage data from all users to learn a scalar reward, higher-granularity approaches must learn more complex patterns (e.g., preference distributions [Siththaranjan *et al.*, 2023] or personalized reward functions [Poddar *et al.*, 2024]), which are more challenging to estimate due to their higher dimensionality and sparser per-pattern data coverage.

A further complication arises from the process of *preference construction* in complex and novel domains, where humans actively construct their preferences rather than simply reveal them [Lichtenstein and Slovic, 2006]. While current reward modeling approaches typically rely on passive data collection and straightforward labeling, research indicates this may be insufficient: the process of *constructing* preferences benefits significantly from deliberate reflection, which helps individuals transform their latent values into concrete preferential choices [Fischhoff, 1991]. This insight is further supported by findings from psychology, consumer research and deliberative polling, which demonstrate that prompting people to actively contemplate their underlying values and reasoning processes leads to more well-defined and consid-

*Our stimuli and prompts are available at https://osf.io/8yxf2/. The behavioral studies in this paper were approved by the University of Waterloo's Human Research Ethics Board (File 46074).

ered preferences [Hauser *et al.*, 2014; Ver Donck *et al.*, 2020; Fishkin and Luskin, 2005].

To address these two challenges, we introduce *Interactive-Reflective Dialogue Alignment* (IRDA), a system that uses large language models (LLMs) to learn personalized reward functions through interactive dialogue. IRDA combines three core components: (1) **reflective language-based preference elicitation** that guides users in articulating their values, (2) **active learning** to strategically select examples for human critique, and (3) **LLM-driven reward modeling** where the LLM directly serves as the reward function by leveraging its in-context learning capabilities to generalize from sparse user feedback. IRDA's architecture directly confronts the preference construction problem by replacing passive labeling with LLM-guided dialogues that provoke deliberate, context-sensitive reflection (System 2 cognition [Fischhoff, 1991]). At the same time, its data-efficient learning strategy mitigates the difficulty of learning personalized reward functions.

We evaluated IRDA with two user studies involving 30 participants in total. The first study (21 participants) focused on building a reward model for each user's personal definition of "respectful behavior," while the second (9 participants) explored ethical decision-making in autonomous vehicles. Across both studies, participants differed widely in their value judgments, and our system was able to capture these individual definitions of value-aligned behavior more accurately than baseline approaches.

Our contributions are as follows:

- A novel pipeline for aligning AI agents to individual values, informed by AI, HCI, and social science.
- A comprehensive evaluation in two distinct domains, demonstrating that our system captures individual preferences more accurately than baselines.
- Empirical characterization of how individuals diverge in their conceptions of value-aligned AI behavior.
- Insights for future research on interactive systems that help end users construct, refine, and operationalize their latent values for both personal and collective alignment.

2 Related Work

Pluralistic Alignment. Recent work has highlighted the importance of moving beyond monolithic reward models and toward approaches that capture heterogeneous or uncertain human preferences in AI alignment. For instance, distributional preference learning (DPL) estimates an entire distribution over possible reward values, thereby accommodating hidden context and diverse annotator criteria [Siththaranjan *et al.*, 2023]. Similarly, methods like MaxMin-RLHF learn a mixture of reward models and optimize an egalitarian objective to avoid disproportionately favoring majority viewpoints [Chakraborty *et al.*, 2024]. Others have proposed user-specific latent variables that personalize reward models without requiring extensive per-user labels [Poddar *et al.*, 2024] or have leveraged meta-learning to reduce feedback requirements [Hejna and Sadigh, 2022]. However, these approaches assume users have direct access to their preferences in novel contexts despite evidence from preference elicitation and psychology literature suggesting otherwise. Our work complements these advances by actively eliciting fine-grained, user-specific preferences and helping users turn their latent values into concrete preferences through guided reflection.

Language-Based Reward Design. A separate line of research explores using large language models (LLMs) to specify or generate reward functions. Some methods prompt LLMs to propose reward code, which is then used to train RL policies via standard optimization [Ma *et al.*, 2024; Xie *et al.*, 2024; Verma *et al.*, 2024; Behari *et al.*, 2024]. Other methods directly treat an LLM as a proxy reward function by prompting it with desired behavior descriptions [Kwon *et al.*, 2022]. These LLM-based approaches have made reward specification more accessible, particularly in domains where handcrafting objectives is difficult. However, they treat preference specification as a one-way street, where users tell the LLM what they want. Our system makes it a two-way dialogue, using LLMs both to help users clarify their preferences and to translate those preferences into reward functions.

Reflection as a Path to Expressing Latent Preferences. Preferences are rarely pre-defined artifacts waiting to be extracted; instead, in new contexts, they form through reflective processes that turns latent values into concrete preferences [Fischhoff, 1991]. In consumer and behavioral research, explicitly prompting users to reflect on trade-offs or alternative perspectives fosters more stable and revealing preference statements [Hauser *et al.*, 2014; Ver Donck *et al.*, 2020]. We build on this insight by weaving reflection into an LLM-based alignment pipeline, enabling users to clarify and externalize their values for AI systems.

Designing Reflection into Dialogue Systems. Within HCI, frameworks like Fleck and Fitzpatrick's [Fleck and Fitzpatrick, 2010] outline how technologies can scaffold reflective thinking, while studies show that structured prompts—scripted or adaptive—support deeper introspection over time [Kocielnik *et al.*, 2018; Wolfbauer *et al.*, 2022; Bentvelzen *et al.*, 2022]. Recent advances in LLM-driven agents further enable flexible "reflective dialogues," guiding users to articulate emergent values and transform them into actionable preferences [Arakawa and Yakura, 2024]. Our contribution is to extend these dialogues to preference elicitation for AI alignment, using LLMs to help users iteratively refine how an AI agent should act.

3 Interactive-Reflective Dialogue Alignment (IRDA) System

We present the *Interactive-Reflective Dialogue Alignment* (IRDA) system, which enables non-expert users to iteratively define a value concept and construct a corresponding reward model for agent training. Our approach is founded on the insight that human values are refined through a process of reflection and iterative feedback [Fischhoff, 1991]. To this end, IRDA employs a dual-loop framework that first elicits user feedback through a *preference construction loop* over a diversity-based pool of trajectories and then refines the model via an *uncertainty reduction loop* over a separate pool.

To begin, let $\mathscr{T}_D$ denote the diversity-based pool of trajectories, where each trajectory $\tau = (s_0, a_0, \ldots, s_T)$ represents the sequence of states and actions executed by the agent. We extract latent features $\phi(\tau) \in \mathbb{R}^d$ from each trajectory and partition $\mathscr{T}_D$ into k clusters $\{\mathscr{C}_1, \ldots, \mathscr{C}_k\}$ using k-means clustering:

$$\{\mathscr{C}_i\}_{i=1}^k = \arg\min_{\{\mathscr{C}_i\}} \sum_{i=1}^{k} \sum_{\tau \in \mathscr{C}_i} \|\phi(\tau) - \mu_i\|_2^2,$$

with cluster centroids

$$\mu_i = \frac{1}{|\mathscr{C}_i|} \sum_{\tau \in \mathscr{C}_i} \phi(\tau).$$

For each cluster, we select a representative trajectory:

$$\tau_i^{cent} = \arg\min_{\tau \in \mathscr{C}_i} \|\phi(\tau) - \mu_i\|_2.$$

In the initial *preference construction loop*, the user specifies a value concept (e.g., "respectfulness") and provides qualitative feedback e_i on each centroid trajectory (e.g., "This is not respectful because..."). Each trajectory is encoded into an ASCII representation $\alpha(\tau)$, and the resulting feedback is aggregated as

$$\mathscr{D}_{fb} = \{(\alpha(\tau_i^{cent}), e_i)\}_{i=1}^k.$$

Once the feedback is collected, an LLM is queried on the entire $\mathscr{D}_{fb}$ to generate a hypothesis about what features the user is using to make their decisions, $\mathscr{H}$, and alternative features the user could consider $\mathscr{A}$:

$$(\mathscr{H}, \mathscr{A}) = G\big(\mathscr{D}_{fb}\big).$$

The user is then asked to respond to the generated hypotheses and alternatives, explaining why these features are or are not significant to their decision-making. This is intended to help the user reflect on their values, which can update their mental model $\mathscr{M}_u$. If $\mathscr{M}_u$ does change, the user returns to the beginning of the *preference construction loop*.

Once the user confirms that $\mathscr{M}_u$ is stable, the system transitions to the *uncertainty reduction loop*. In this stage, we consider a separate uncertainty-based pool of trajectories, $\mathscr{T}_U$. We iteratively refine our reward model on this pool. The reward model is based on an LLM that is prompted to assess whether a trajectory is aligned or not. The LLM is given the entire conversation history $\mathscr{C}$ (which includes all user feedback, system prompts, and responses), the encoded trajectory $\alpha(\tau)$, and environment details such as symbol meanings and action spaces (EnvDesc) and outputs token probabilities for labels such as "respectful" and "disrespectful." Specifically, the alignment probability $p_\theta(1|\tau)$ and misalignment probability $p_\theta(1|\tau)$ are computed as

$$p_\theta(1|\tau), p_\theta(0|\tau) = f_{\text{LLM}}\big(\text{EnvDesc}, \mathscr{C}, \alpha(\tau)\big),$$

where $p_\theta(1|\tau)$ is the token probability for the "aligned" token (e.g., "respectful") and $p_\theta(0|\tau)$ is the token probability for the "misaligned" token (e.g., "disrespectful").

The associated uncertainty is defined by

$$U(\tau) = 1 - \big|p_\theta(1|\tau) - p_\theta(0|\tau)\big|,$$

Algorithm 1 *Interactive-Reflective Dialogue Alignment*

Input: $\mathscr{T}_D$, $\mathscr{T}_U$, value v, threshold ε, EnvDesc
Preprocessing: For each $\tau \in \mathscr{T}_D$, extract $\phi(\tau)$; cluster via k-means; select representatives $\{\tau_i^{cent}\}$.
repeat *Preference Construction Loop*
  For each τ_i^{cent}, obtain/update label e_i; form

$$\mathscr{D}_{fb} = \{(\alpha(\tau_i^{cent}), e_i)\}.$$

  Query the LLM with $\mathscr{D}_{fb}$ to yield feature hypotheses $\mathscr{H}$ and alternatives $\mathscr{A}$.
  User responds to $(\mathscr{H}, \mathscr{A})$ and refines $\mathscr{M}_u$.
until User confirms $\mathscr{M}_u$ is stable
repeat *Uncertainty Reduction Loop*
  for each $\tau \in \mathscr{T}_U$ **do**
    $p_\theta(1|\tau), p_\theta(0|\tau) = f_{\text{LLM}}(\text{EnvDesc}, \mathscr{C}, \alpha(\tau))$
    $U(\tau) = 1 - |p_\theta(1|\tau) - p_\theta(0|\tau)|$
  end for
  $\tau^* = \arg\max_{\tau \in \mathscr{T}_U} U(\tau)$.
  Query the user for e^* on τ^*; update

$$\mathscr{D}_{fb} \leftarrow \mathscr{D}_{fb} \cup \{(\alpha(\tau^*), e^*)\}.$$

until $U(\tau^*) < \varepsilon$
Final Reward Model: For any new τ, define

$$R(\tau) = \mathbb{I}\big[p_\theta(1|\tau) > p_\theta(0|\tau)\big],$$

Return: $R(\cdot)$.

and the trajectory with maximum uncertainty is selected,

$$\tau^* = \arg\max_{\tau \in \mathscr{T}_U} U(\tau),$$

and the user is queried for an explanation e^* regarding τ^*. This new feedback $(\alpha(\tau^*), e^*)$ is appended to $\mathscr{D}_{fb}$, and the uncertainty reduction loop is repeated until $U(\tau^*) < \varepsilon$ for a set threshold ε.

The final reward model is an LLM that, using the complete conversation history $\mathscr{C}$ and all accumulated feedback $\mathscr{D}_{fb}$, classifies new trajectories as aligned or misaligned with the user's value. It outputs a binary decision based on token probabilities:

$$R(\tau) = \mathbb{I}\big[p_\theta(1|\tau) > p_\theta(0|\tau)\big].$$

This unified process—beginning with diversity-based sampling, followed by preference construction, and culminating in uncertainty-driven refinement yields a final LLM-based reward model. This final LLM-based reward model leverages the entire conversation history and all explained examples. The full process is formalized in Algorithm 1.

4 Study Design & Methodology

We evaluated our system in two studies: Study 1 investigates the utility of our system for learning about participants' definition of *respectful* agent behavior. Study 2 investigates the utility of our system for learning about participants' decision-making in moral dilemmas involving an agent (autonomous

vehicle). Our studies employ a within-subject design, collecting data from each participant to train and test each method. Our studies aim to answer the following three questions:

- **RQ1:** How do individuals' interpretations of value-aligned AI behavior differ?
- **RQ2:** Does structured reflection enhance language-based reward modeling?
- **RQ3:** When is individualized language-based reward modeling effective?

4.1 Environments

Multi-Agent Apple Farming Environment (Study 1). A 6×6 grid contains apples and garbage, with one "main" (blue) agent and three "background" (grey) agents. Each agent "owns" one of four 3×3 orchards. Two background agents remain stationary, and one moves freely. The main agent is rewarded by picking apples (none for collecting garbage). Participants assess whether the blue agent agent behaves "respectfully."

Moral Machine Environment (Study 2). Adapted from [Awad *et al.*, 2018], this environment features ethical dilemmas in which an autonomous vehicle must stay on course or swerve, potentially harming different combinations of pedestrians or passengers (including children, adults, the elderly, and animals). Each outcome varies in factors like legality, social status, and species. Participants decide which outcome the car should choose.

4.2 Participants

In Study 1, we recruited 21 participants from the University of Waterloo (18 to 39 age range, M=23.86, 7 self-identified as male and 14 as female). When asked to rate their level of familiarity with reinforcement learning on a 5-point Likert ranging from "very unfamiliar" (1) to "very familiar" (5), the mean level of familiarity was 2.48, with the mode and median being 2. he Likert-scale data for Study 1 (Figure 1, top bar) highlights that more than half of participants were "unfamiliar" or "very unfamiliar" with reinforcement learning.

In Study 2, we recruited 9 participants from the University of Waterloo (18 to 33 age range, M=25.66, 6 self-identified as male and 3 as female). When asked to rate their level of familiarity with reinforcement learning on a 5-point Likert ranging from "very unfamiliar" to "very familiar," the mean level of familiarity was 3.55, with the mode and median being 3. The Likert-scale data for Study 2 is visualized in Figure 1 (bottom).

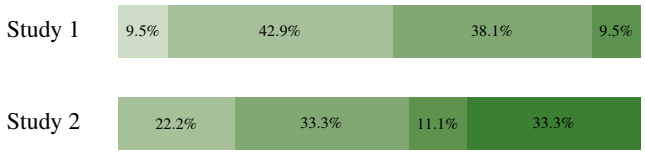

Figure 1: Participant familiarity with reinforcement learning for Study 1 (top) and Study 2 (bottom). Key: Very Unfamiliar; Unfamiliar; Neutral; Familiar; Very Familiar.

4.3 Procedure

Participants used our system to specify how they would like the agent to act. Participants then labeled 50 scenarios and were interviewed.

Introduction (∼5min) - After participants completed the consent form and demographic questions, we thoroughly explained the environment mechanics so that differences reflected their opinions rather than assumptions about the setup.

Dialogue - The participant began by conversing with the system about the agent's behavior following the process described in section 3. To control the amount of time users spent, we limited the user to one preference construction loop and one uncertainty reduction loop.

Labeling - Following the participants' dialogue interaction with the system, participants labeled 20 scenarios. Each participant labeled the same scenarios, which allowed us to assess how much the participants agreed on the labels.

Semi-structured Interview (∼10min) - After completing the labeling task, participants were asked about their ability to communicate their decision-making to the system, including their capacity to articulate label choices, any difficulties in decision-making, and potential changes in their labeling behavior over time.

4.4 Baseline Comparisons

We compared our system to several baselines, including another language-based system and various supervised learning approaches.

Language-Based Baseline (L^B)

Kwon et al. [2022] proposed a reward modeling pipeline for text-based environments where the user selects multiple examples from a palette of examples of the agent behaving as they would desire, accompanied by explanations. We modify their pipeline in the following way: Instead of asking the user to select examples from a handcrafted palette, we choose the examples the user sees with the diversity- and uncertainty-based sampling procedures described and in section 3. This system differs from IRDA in that it does not engage the user in reflective dialogue.

Supervised Learning Baselines

We evaluated our approach against neural network-based supervised learning methods, including both individual models per participant and collective models trained on aggregated data. In Study 1, we implemented multi-layer perceptron models (one hidden layer, 32 neurons) using a tensor encoding of the trajectory: individual models ($\text{MLP}_i^{\text{ind}}$) for each participant i and a collective model (MLP^{col}) trained on all participant data. Study 2 expanded this comparison to include both MLP models with 26-dimensional Moral Machine scenario vectors and convolutional neural networks ($\text{CNN}_i^{\text{ind}}$ and CNN^{col}) using scenario image inputs. The CNNs used two convolutional layers (16 and 32 filters) with max pooling, followed by fully connected layers reducing to 64 dimensions. Each supervised learning model was incrementally trained with up to 30 examples *per participant* using the Adam optimizer (learning rate 0.001).

Table 1: Mapping of Analysis Methods to Research Questions

Analysis Method	RQ1	RQ2	RQ3
Inter-Annotator Agreement	✓		✓
Evaluation of Language-Based Reward Model Performance		✓	
Comparison to Supervised Learning	✓		✓
Qualitative Analysis of Participant Decision Making	✓		
Analysis of Feature Similarity Between Participants	✓		✓
Interview Thematic Analysis		✓	✓

4.5 Analysis

To answer our research questions, we employ a mixed-methods approach, combining quantitative analyses of model performance and inter-annotator agreement with qualitative analyses of participant decision-making processes and experiences.

Inter-Annotator Agreement
We assess the inter-annotator agreement between participants on the test set of scenarios they labelled in each study. Since each participant labelled the same test scenarios, we can use Fleiss' kappa value to quantify the inter-annotator agreement between the participants [Landis and Koch, 1977]. Generally, kappa statistics below 0 indicate "poor" agreement and kappa statistics above 0.8 indicate "nearly perfect" agreement [Landis and Koch, 1977].

Evaluation of Language-Based Reward Model Performance
We evaluated our system against a baseline without dialogic reflection using a performance metric P, where P_i^{IRDA} and P_i^B represent participant i's metrics for our system and baseline, respectively. Study 1 used balanced accuracy due to class imbalance, while Study 2 used accuracy. For each participant, both systems generated rewards for 20 non-training scenarios, yielding 20 pairs of P values. We conducted three statistical tests on the P values:

1. We bootstrapped 95% confidence intervals for the mean by resampling 10,000 times with replacement.
2. For each participant, we calculated the difference $\Delta P_i = P_i^{\text{IRDA}} - P_i^B$ and bootstrapped these differences in the same way.
3. P values were compared using the Wilcoxon signed-rank test, chosen for its robustness to non-normal distributions and reduced false positives [Bridge and Sawilowsky, 1999].

Comparison to Supervised Learning
We compared our language-based systems to traditional supervised learning approaches. Both the individual models ($\text{MLP}_i^{\text{ind}}$ and $\text{CNN}_i^{\text{ind}}$) and the collective models (MLP^{col} and CNN^{col}) were trained incrementally, gradually increasing the number of samples used per participant. This methodology allowed us to analyze how model performance evolved with increasing data availability. For each increment, we calculated P_i^{ind} and P_i^{col} for each participant i. To ensure robustness, we bootstrapped these values with replacement using 10,000 resamples.

Qualitative Analysis of Participant Decision Making
We conducted a detailed analysis of the message exchanges between participants and the system to gain insight into participants' decision-making processes. We employed an inductive coding approach, systematically reviewing the messages to identify key features and criteria that participants used in their decision-making. Our coding process involved multiple passes through the data, with iterative refinement of the codebook to ensure it captured the full range of decision-making strategies observed.

Analysis of Feature Similarity Between Participants
To quantify how similar participants were in their use of decision-making features, we employed the Jaccard similarity coefficient. This measure calculates the overlap between two sets of items, which, in our case, are features the two participants used to make decisions [Jaccard, 1912]. We computed the Jaccard similarity coefficient for every pair of participants, using the set of decision-making features each participant employed (as identified in our qualitative analysis). To estimate the overall similarity across our participant pool, we then calculated the mean of these pairwise Jaccard coefficients. We used bootstrapping with 10,000 resamples to determine the 95% confidence intervals.

Thematic Analysis of Interview Data
We conducted semi-structured interviews with participants to understand their experiences. The interview transcripts were analyzed using a thematic analysis approach guided by the principles outlined by Braun and Clarke [2006]. We followed a six-phase process: familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report.

5 Results: Study 1 - Multi-Agent Apple Farming

On average, participants took 15 minutes 57 seconds (SD = 6 min. 43 sec., range: 6 min. 59 sec. - 30 min. 55 sec.) to complete the dialogue with the system and 13 minutes 37 seconds (SD = 3 min. 2 sec., range: 6 min. 55 sec. - 18 min. 26 sec.) to complete the labelling of 50 trajectories. Of 21 participants, 7 ($33.\bar{3}\%$) entered the *preference construction loop* for one iteration.

S1 - Inter-Annotator Agreement
We observed a Fleiss' kappa value between all participants' labels on the 50 labelled trajectories of $\kappa = 0.336$, indicating "fair" agreement among participants [Landis and Koch, 1977]. The Fleiss' kappa statistic of 0.336 we observed lends credence to the idea that human values and preferences are subjective and personal.

S1 – Evaluation of Language-Based Reward Model Performance.
On average, the reward models produced by our pipeline (IRDA) received significantly higher balanced accuracy

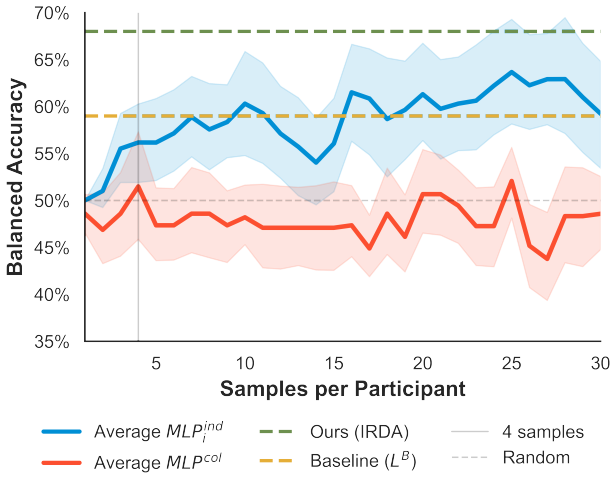

Figure 2: Balanced accuracy of models vs. samples per participant in **Study 1**. Blue line shows average individual MLP (MLP^{ind}); red shows collective MLP (MLP^{col}). Our IRDA system (green dashed) and baseline (L^B, yellow dashed) used 4 samples per participant (vertical gray line). Collective model trained on 21x samples shown (21 participants). Shaded areas: 95% confidence intervals. Gray dashed line: random performance.

scores (measured in percentages) than the baseline system (L^B) by 9% (95% CI: [5%, 13%], M = 68% vs. M = 59%, p=.002). This indicates that structured reflection is beneficial.

S1 – Comparison to Supervised Learning

With all 30 training samples, the average balanced accuracy of the individual models (MLP_i^{ind}) was 59% (95% CI: [53%, 65%]) while the collective model (MLP^{col}) achieved 48% (95% CI: [46%, 50%]). This indicates that participant value definitions varied significantly. Figure 2 illustrates the relationship between model performance and the number of samples provided per participant.

S1 – Qualitative Analysis of Participant Decision Making

Although our system can align AI agents with various values, we focused on respect to examine how individuals interpret even a single value differently. Analysis of participant conversations revealed 12 distinct behavioral features used to evaluate respectful agent behavior. Usage varied significantly - P1 relied solely on whether agents stayed in their quadrant, while P5, P7, and P10 each employed seven features. Only one participant pair shared identical feature sets, and most participants combined them using hierarchical and conditional rules. While the agent staying in its quadrant was the most common feature, features varied in temporal scope - from static properties (current quadrant location) to multi-step sequences (collecting garbage before apples).

S1 – Analysis of Feature Similarity Between Participants.

We observed an average Jaccard similarity coefficient between all pairs of participants' feature usage of $J = 0.357$, with a 95% confidence interval of $(0.333, 0.3813)$.

S1 – Thematic Analysis of Interview Data

Our thematic analysis revealed two main themes: participants' evolving definitions of respect and the system's impact on this evolution.

Evolving Definitions of Respect. Participants' understanding of respect developed significantly through system interaction. Initial definitions focused on simple concepts like spatial boundaries and task-specific behaviors (P4, P6, P7, P10, P19). Through engagement with examples and system feedback, these views became more nuanced. Participants who encountered examples challenging their initial perspectives often expanded their conceptualization of respect (P3, P8, P10, P13, P18, P19, P20). This evolution aligns with consumer research showing that reflection and realistic decision-making improve preference reporting [Hauser *et al.*, 2014].

Specific Impact of System Hypothesis. The system's hypotheses and alternative features actively shaped participants' respect definitions. The presentation of alternatives prompted reevaluation and revision of initial concepts (P3, P8, P13, P18, P20). While some participants maintained their original views (e.g., P19), the system's suggestions helped others refine their understanding (P13, P18) or consider new perspectives (P3, P20), demonstrating the value of reflective dialogue.

6 Results: Study 2 - The Moral Machine

On average, participants took 18 minutes 28 seconds (SD = 7 min. 15 sec., range: 12 min. 0 sec. - 34 min. 34 sec.) to complete the dialogue with the system and 11 minutes 51 seconds (SD = 4 min. 15 sec., range: 4 min. 46 sec. - 17 min. 04 sec.) to complete the labelling of 50 trajectories. Of 9 participants, 1 (11.11%) entered the *preference construction loop* for one iteration.

S2 - Inter-Annotator Agreement

We observed a Fleiss' kappa value between all participants' labels on the 50 labeled trajectories of $\kappa = 0.460$, indicating "moderate" (higher than "fair") agreement among participants [Landis and Koch, 1977].

S2 – Evaluation of Language-Based Reward Model Performance.

On average, the reward models produced by our pipeline (IRDA) received significantly higher accuracy scores (measured in percentages) than the baseline system (L^B) by 12% (95% CI: [4%, 27%], M = 65% vs. M = 53%, p=.05). This adds more evidence in favor of the effectiveness of structured reflection.

S2 – Comparison to Supervised Learning

With 30 training samples, individual MLPs achieved 79% accuracy (95% CI: [74%, 84%]) while the collective MLP reached 77% (95% CI: [75%, 78%]). For CNNs, individual models achieved 67% accuracy (95% CI: [61%, 73%]) while the collective model (CNN^{col}) reached 77% (95% CI: [70%, 83%]). Figure 3 shows performance versus sample count. These results suggest that collective methods excel with high participant agreement, particularly for complex

learning problems like CNN-based image processing, where sample pooling proves beneficial.

S2 – Qualitative Analysis of Participant Decision Making By analyzing the participants' conversations with our system, we identified nine features they used in their decision-making in Study 2. The most common features were minimizing casualties (8/9 participants) and traffic rule compliance (8/9 participants). Most participants combined features conditionally and hierarchically.

S2 – Analysis of Feature Similarity Between Participants. We observed an average Jaccard similarity coefficient between all pairs of participants' feature usage of $J = 0.464$, with a 95% confidence interval of $(0.403, 0.526)$.

S2 – Thematic Analysis of Interview Data Through our thematic analysis, we found two main themes: (1) participants' definitions of ethical decision making evolved throughout the activity, and (2) participants' decisions were primarily based on explicit reasoning but sometimes relied on intuition.

Decision-making Evolution. Participants' decision-making evolved differently during the study. Some expanded their criteria as they encountered more scenarios (P3: "realized the need to consider new factors when initial factors were equal"), reinforcing Study 1's findings about preference evolution. Others maintained consistent frameworks throughout (P2: "rules remained consistent").

Intuition vs. Explicit Reasoning. While most participants could articulate clear reasoning, some relied on intuition for complex scenarios. P7 reported using "first instinct" or "vibes" for several challenging cases.

7 Discussion

Our evaluation of our system reveals important insights about value alignment in AI systems. Our results suggest significant value diversity exists in the context of AI alignment and that language-based reward modeling with LLM-guided reflection can be an effective method for learning individualized reward models that preserve this value diversity.

Value Diversity (RQ1). In Study 1, focusing on respectful behavior, we observed substantial diversity in how participants defined and evaluated respect. The low inter-annotator agreement ($\kappa = 0.23$) and feature similarity ($J = 0.357$) suggest fundamental differences in value interpretation rather than noise. This was further evidenced by the stark performance gap between individual and collective models - while individual models achieved 59% accuracy with just 30 samples, collective models failed to surpass random performance despite access to 21 times more data.

Study 2, examining ethical decisions in autonomous vehicles, revealed more homogeneous preferences ($\kappa = 0.460$, $J = 0.464$). Here, collective models outperformed individual approaches. This contrast between studies highlights how value diversity varies by context and challenges the assumption that universal values can be embedded in AI systems.

The Effectiveness of Reflection (RQ2). Our reflection-based approach proved effective across both contexts, outperforming baseline methods even when participants held

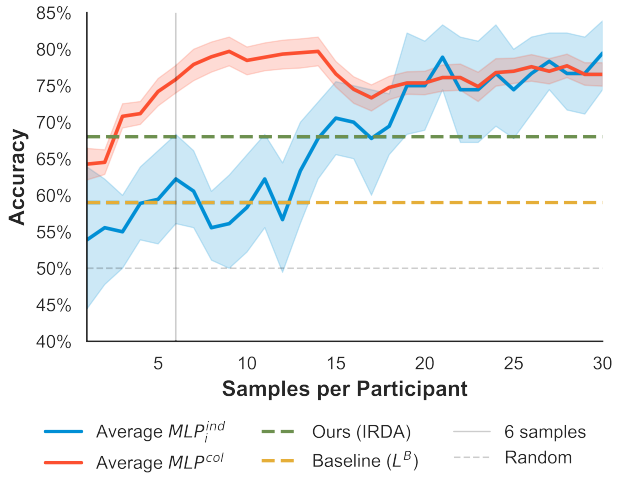

Figure 3: Model accuracies vs. samples per participant in **Study 2**. Blue: average individual MLP (MLP^{ind}); red: collective MLP (MLP^{col}); green dashed: our IRDA approach; yellow dashed: baseline L^B. Shaded areas show confidence intervals. Vertical gray line: 6-sample training point for IRDA and L^B. Collective model used 9x samples shown (9 participants). Gray dashed: random performance.

fixed opinions. This suggests that structured reflection enhances preference communication regardless of preference malleability, aligning with dual-process theories in cognitive psychology that suggest reflection can enhance articulation [Evans, 2019]. Given this, we suggest effective preference elicitation requires more sophisticated interaction paradigms than simple feedback collection and should engage users in reflection.

Contextual Efficacy of Language-Based RMs (RQ3). The effectiveness of our approach varies with context: it excels with heterogeneous preferences and limited samples but becomes less critical when preferences are homogeneous and data is abundant. This suggests a complementary approach to value alignment, where methods are selected based on preference heterogeneity and data availability.

Limitations and Future Work. Our participant pool was relatively homogeneous, so results may not generalize to broader populations or values. Additionally, repeatedly querying an LLM at training time can become prohibitively expensive, limiting scalability for large-scale RL tasks. An alternative is having the LLM produce a standalone code-based reward function, which is more computationally efficient but less flexible. Future work should evaluate this approach in more diverse settings (including LLM alignment tasks) and develop principled methods for aggregating individualized reward models. Social choice offers one promising path for navigating these aggregation decisions [Conitzer *et al.*, 2024]. In sum, our findings highlight the importance of reflection-based elicitation for capturing preferences and underscore the need for methods designed to accommodate genuine value pluralism in AI alignment.

References

[Arakawa and Yakura, 2024] Riku Arakawa and Hiromu Yakura. Coaching copilot: blended form of an llm-powered chatbot and a human coach to effectively support self-reflection for leadership growth. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–14, 2024.

[Awad *et al.*, 2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[Bai *et al.*, 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. arXiv:2204.05862 [cs].

[Behari *et al.*, 2024] Nikhil Behari, Edwin Zhang, Yunfan Zhao, Aparna Taneja, Dheeraj Nagaraj, and Milind Tambe. A decision-language model (dlm) for dynamic restless multi-armed bandit tasks in public health. *arXiv preprint arXiv:2402.14807*, 2024.

[Bentvelzen *et al.*, 2022] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):2:1–2:27, March 2022.

[Braun and Clarke, 2006] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[Bridge and Sawilowsky, 1999] Patrick D Bridge and Shlomo S Sawilowsky. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology*, 52(3):229–235, 1999.

[Casper *et al.*, 2023] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[Chakraborty *et al.*, 2024] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

[Conitzer *et al.*, 2024] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9346–9360. PMLR, 21–27 Jul 2024.

[Evans, 2019] Jonathan St BT Evans. Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, 25(4):383–415, 2019.

[Fischhoff, 1991] Baruch Fischhoff. Value elicitation: Is there anything in there? *American Psychologist*, 46(8):835, 1991.

[Fishkin and Luskin, 2005] James S Fishkin and Robert C Luskin. Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta politica*, 40:284–298, 2005.

[Fleck and Fitzpatrick, 2010] Rowanne Fleck and Geraldine Fitzpatrick. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, pages 216–223, 2010.

[Friedman *et al.*, 2013] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. *Early Engagement and New Technologies: Opening Up the Laboratory*, pages 55–95, 2013.

[Hauser *et al.*, 2014] John R Hauser, Songting Dong, and Min Ding. Self-reflection and articulated consumer preferences. *Journal of Product Innovation Management*, 31(1):17–32, 2014.

[Hejna and Sadigh, 2022] Joseph Hejna and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop RL. In *6th Annual Conference on Robot Learning*, 2022.

[Jaccard, 1912] Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[Kocielnik *et al.*, 2018] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):70:1–70:26, July 2018.

[Kwon *et al.*, 2022] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[Landis and Koch, 1977] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[Le Dantec *et al.*, 2009] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1141–1150, 2009.

[Lichtenstein and Slovic, 2006] Sarah Lichtenstein and Paul Slovic. The construction of preference: An overview. *The construction of preference*, 1:1–40, 2006.

[Ma *et al.*, 2024] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Poddar *et al.*, 2024] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Schwartz, 1992] Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology/Academic Press*, 1992.

[Siththaranjan *et al.*, 2023] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Understanding hidden context in preference learning: Consequences for rlhf. In *The Twelfth International Conference on Learning Representations*, 2023.

[Ver Donck *et al.*, 2020] Niki Ver Donck, Geert Vander Stichele, Isabelle Huys, et al. Improving patient preference elicitation by applying concepts from the consumer research field: narrative literature review. *Interactive Journal of Medical Research*, 9(1):e13684, 2020.

[Verma *et al.*, 2024] Shresth Verma, Niclas Boehmer, Lingkai Kong, and Milind Tambe. Balancing act: Prioritization strategies for LLM-designed restless bandit rewards. In *Workshop on Socially Responsible Language Modelling Research*, 2024.

[Wolfbauer *et al.*, 2022] Irmtraud Wolfbauer, Viktoria Pammer-Schindler, Katharina Maitz, and Carolyn P. Rose. A Script for Conversational Reflection Guidance: A Field Study on Developing Reflection Competence With Apprentices. *IEEE Transactions on Learning Technologies*, 15(5):554–566, October 2022.

[Xie *et al.*, 2024] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.