Democratizing Reward Design for Personal and Representative Value-Alignment

Carter Blair
University of Waterloo
Waterloo, Canada
cblair@uwaterloo.ca

Kate Larson
University of Waterloo
Waterloo, Canada
kate.larson@uwaterloo.ca

Edith Law
University of Waterloo
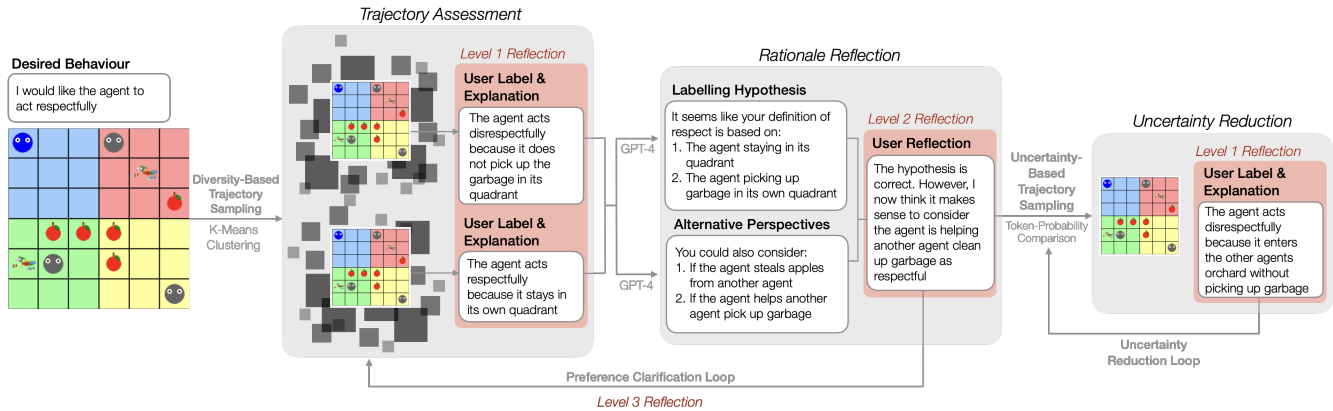Waterloo, Canada
edith.law@uwaterloo.ca

Figure 1: Dialogue-Based Preference Elicitation Portion of our Reward Design System.

ABSTRACT

Aligning AI agents with human values is challenging due to diverse and subjective notions of values. Standard alignment methods often aggregate crowd feedback, which can result in the suppression of unique or minority preferences. We introduce Interactive-Reflective Dialogue Alignment, a method that iteratively engages users in reflecting on and specifying their subjective value definitions. This system learns individual value definitions through language-model-based preference elicitation and constructs personalized reward models that can be used to align AI behaviour. We evaluated our system through two studies with 30 participants, one focusing on "respect" and the other on ethical decision-making in autonomous vehicles. Our findings demonstrate diverse definitions of value-aligned behaviour and show that our system can accurately capture each person's unique understanding. This approach enables personalized alignment and can inform more representative and interpretable collective alignment strategies.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Active learning settings**; *Reinforcement learning*; **Learning from critiques**;

KEYWORDS

Reinforcement Learning from Human Feedback, Democratization, Personalization, Value-Alignment

1 INTRODUCTION

As AI agents take on more tasks and affect personal parts of our lives, it is increasingly important to align their behaviour with human values at both individual and collective levels. This alignment challenge spans from personalized interactions, like an AI agent's respectful behaviour in a household, to broader societal collective decisions.

Current AI alignment approaches, such as reinforcement learning from human feedback (RLHF), often rely on feedback aggregated from many users [6, 15]. This aggregation implicitly makes collective decisions about AI behaviour, potentially marginalizing minority viewpoints or unique preferences [68]. By "averaging out" personal differences, these methods risk creating agents that ignore the values of minority groups or those with uncommon preferences. Accurate individual reward models offer a solution to this problem. These models enable personalization by tailoring AI behaviour to each user's unique values and beliefs when appropriate while simultaneously providing a foundation for more representative and interpretable collective decision-making. By understanding individual stakeholder preferences, we can aggregate diverse viewpoints more fairly and make informed trade-offs in group contexts. This

approach addresses both the need for personalized AI interactions and the challenge of making ethical collective decisions.

However, creating personalized reward models presents several challenges. First, finding the most effective way for humans to convey their goals and desires to AI systems remains an open question [15]. Second, end users of AI systems may lack the technical skills or necessary training to provide feedback to an agent [7]. Third, it is impractical to expect individuals to teach an agent how to behave when many examples may be required.

In this work, we present *Interactive-Reflective Dialogue Alignment*, an interactive system for aligning AI agents to individual values that is novice-friendly and sample-efficient. The user-facing side is a simple chat interface that prompts users to explain their desired behavioural patterns and selectively sends examples of the agent's behaviour to solicit feedback using active learning techniques [49]. The system's textual messages are designed to encourage users to reflect deeply on their value definitions, inspired by prior work on designing for reflection [9, 25, 34, 73]. Using the feedback gathered from the dialogue, we create a language-based reward model that leverages the in-context learning abilities of large language models [14].

We evaluated *Interactive-Reflective Dialogue Alignment* through two studies involving a total of 30 participants. In the first study, 21 participants used the system to build a reward model for their personal definitions of respectful behaviour. The second study involved 9 participants and focused on ethical decision-making in autonomous vehicles. We found that participants had widely varying definitions of value-aligned behaviour across both studies and that our system could capture these subjective and personal definitions significantly more accurately than baseline systems.

Our contributions are as follows:

- A novel, accessible, and theoretically grounded pipeline for aligning AI agents to individual values and preferences, drawing upon insights from AI, HCI, and social science research.
- A comprehensive evaluation of the system across two distinct domains, demonstrating its ability to capture individual human values and ethical preferences.
- Empirical evidence highlighting the diversity of individual interpretations of value-aligned behaviour.
- Insights for future work on enabling end users to interactively align AI agents with their personal values, including potential applications in both individual and collective alignment contexts.

2 BACKGROUND AND RELATED WORK

Our work intersects human-computer interaction, reinforcement learning, and value alignment. We draw on insights from AI, HCI, and social science research to develop a system that accurately captures individual values in a reward model. Below, we outline the related work most relevant to our system and provide the necessary background to understand it.

2.1 Human Values and Fuzzy Preferences

Values are principles that guide human behaviour and ethical judgments [66]. Friedman et al. describe values as "what a person or group of people consider important in life" [27]. In the context of

AI systems, values can manifest in various ways. For example, in a household setting, the value of respect might manifest as an AI assistant using appropriate language or maintaining privacy. In the context of autonomous vehicles, values might include prioritizing passenger safety while also considering the welfare of pedestrians and other road users. Importantly, values vary across cultures, individuals, and contexts [40].

While values themselves may be clearly stated, their interpretation and application often become fuzzy when considered in specific contexts [70]. This ambiguity is further compounded when attempting to translate these human values into computational terms for AI alignment [64]. Users may struggle to articulate their values or determine which AI behaviours align with their values.

The concept of fuzzy preferences [62] provides insight into this challenge. Fuzzy decision-making acknowledges that preferences, attributes, and objectives in decision processes can be imprecise. In AI alignment, this fuzziness manifests in the difficulty of precisely defining and measuring value alignment. Users may not have clear, *a priori* knowledge of what constitutes aligned behaviour, and the importance of different behavioural features may be unclear or change based on context.

Our work addresses these challenges by developing a system that helps users articulate and refine their fuzzy value definitions, enabling more precise alignment of AI systems with individual human values.

2.2 Designing for Reflection

Reflection has several notable benefits that can help users clarify their understanding of their preferences and values. For example, it has been found that prompting medical practitioners to reflect can increase diagnostic accuracy [19, 23, 45]. Moreover, in consumer research, it has been found that prompting consumers to engage in preference self-reflection can lead to more accurate reporting of preferences and that engaging consumers in realistic decisions can increase preference elicitation accuracy [31].

Designing for reflection has been of interest to HCI researchers for some time [28, 42, 56, 63, 76]. Fleck & Fitzpatrick propose a framework for designing for reflection that includes definitions of various levels of reflection and techniques for supporting reflection [25]. They define five levels of reflection, ranging from simply revisiting past events to critical reflection that considers wider social or moral contexts. Levels 1-3 are most relevant to our work as they inspired the design of our system: Level 1 involves explaining or justifying actions/events. Level 2 involves exploring relationships and considering different perspectives and hypotheses. Level 3 involves fundamentally challenging assumptions and transforming one's understanding or practice.

LLMs present an opportunity for more tailored and dynamic dialogue to engage users in reflection. Some work has explored the use of LLMs for generating reflective prompts in a design template [75], while others have used language prompts to engage users in reflection through single questions or sentence starters [24, 32, 61, 72]. Further extensions of this approach have included the use of pre-scripted, multi-message dialogues to facilitate reflection [35, 74].

We extend this body of work by integrating reflective dialogue techniques with large language models, creating a dynamic and

personalized reflection process that helps users clarify their values for AI alignment.

2.3 Reinforcement Learning and Its Variants

Reinforcement Learning (RL) is a branch of machine learning in which an agent learns a policy, i.e., a strategy for selecting actions in an environment, to maximize the rewards it receives. Traditionally, the agent designer crafts a reward function to incentivize desirable behaviour. However, reward design is notoriously challenging and can result in reward hacking where an agent learns behaviour that gets high reward, but that runs counter to the designers intent [4, 20, 22].

Inverse reinforcement learning (IRL) was developed in response to difficulties with reward design. In IRL, the goal is to learn a reward function from expert demonstrations [1, 50, 51, 59, 77]. This can be useful when a desired behaviour is easy to demonstrate but difficult to design a reward function for. However, since multiple reward functions can often explain the observed behaviour, the reward function lacks interpretability and is inherently ambiguous [46].

Reinforcement learning from human feedback (RLHF) instead learns a reward model from human feedback [6, 18, 78]. This approach is especially useful when aligning agents to human preferences as the agent can optimize for getting "good" feedback. There are three main parts of RLHF: Feedback collection where humans provide feedback, reward modelling where the feedback is turned into a reward model, and policy optimization where the reward model is used to train a policy/agent with reinforcement learning [15].

Our approach bridges the gap between RL and HCI by combining RLHF techniques with user-friendly interfaces and reflective dialogue, making the process of aligning AI agents with individual values more accessible to non-expert users.

2.4 Preference Elicitation and Active Learning

When using language feedback for RLHF, it is natural to implement dialogue-based preference elicitation as the mechanism for collecting feedback. In dialogue-based preference elicitation, the goal is to understand a user's preferences through the use of dialogue [57, 58]. There are two main approaches to preference elicitation: item-based preference elicitation, where the user is asked about their preferences regarding specific things [2, 17, 30, 44, 65] and feature-based preference elicitation where the user is asked about general features or attributes [43, 69, 71]. Some methods use both item- and feature-based methods [11], similar to our approach. However, previous work has focused on content recommendation, not agent alignment.

When performing item-based preference elicitation, we must select items to query the user about. Active learning is a subfield of machine learning where the learner is responsible for choosing which examples to request labels for [60, 67]. Being selective about which examples to query the user about is important in cases where labels are expensive to obtain [54]. There are generally three strategies for sampling items to query the user about, namely random-, diversity-, and uncertainty-based sampling [48]. Random sampling selects items arbitrarily, while diversity-based sampling chooses

items that are most different from each other. Uncertainty-based sampling, in contrast, focuses on items the model is least confident about [49].

Our system integrates these approaches into an interactive dialogue framework, combining item- and feature-based preference elicitation with active learning and reflective dialogue to capture individual value definitions efficiently.

3 INTERACTIVE-REFLECTIVE DIALOGUE ALIGNMENT (IRDA) SYSTEM

We introduce an interactive system, named *Interactive-Reflective Dialogue Alignment* (IRDA), which aims to enable end users with no particular expertise in machine learning to define agent behaviour that is aligned with their subjective definitions of a human value and generate a reward model that can be later used to train an agent based on this definition. We will begin by outlining our design aims, providing a high-level overview of the system and user flow, and then describe each component in more detail.
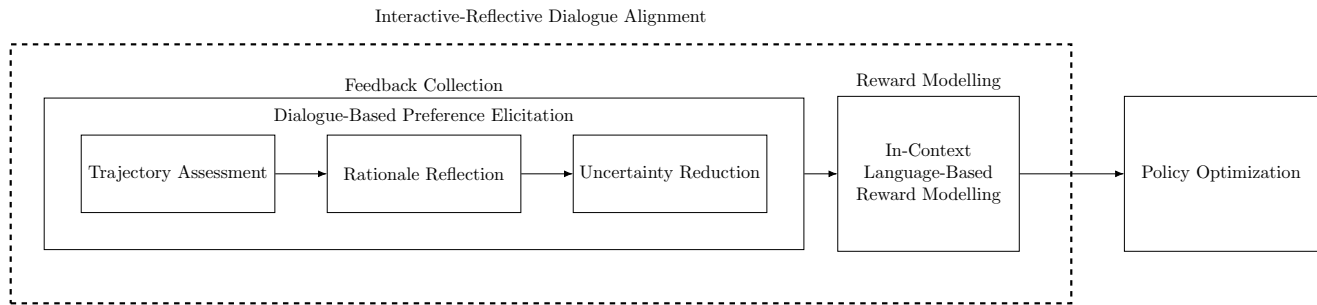
3.1 Design Aims

Our system is intended for the end-users of AI agents, and we do not assume that the users of our system will have any technical expertise. As such, the design aims for our system are centred around the idea of making the system easy and quick to use while maintaining good performance. In particular, we require:

(1) The system should not require any special technical knowledge (e.g., how to program or how to design a "good" reward function).
(2) The system should be sample-efficient (i.e., strategically selecting the most informative examples to ask users for feedback on).
(3) The system should be able to capture individual differences and unique conceptions of the appropriate behaviour associated with a value.
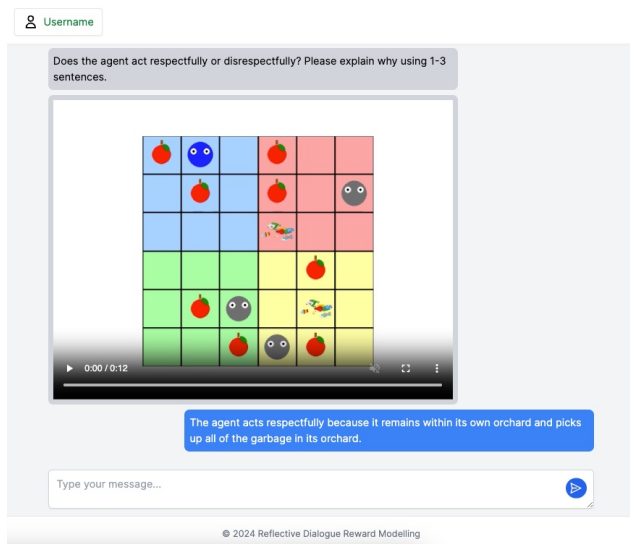
3.2 System and User Flow

Upon entering our system, users are presented with a greeting message outlining the purpose of the system as well as a preview of the environment. To better illustrate our system, we will use one of the environments used in our evaluation, the *Multi-Agent Apple Farming Environment*, as the running example. The environment is a 6×6 grid where the agents are rewarded for picking apples and receive no reward for collecting garbage. There is one blue "main" agent in the environment and three grey "background" agents as shown in Figure 1. Each one of the agents owns one of the four 3×3 quadrants, which each represent an orchard. Each agent is free to move around the whole grid; however, two of the three background agents are programmed to be stationary. The main agent is the agent whose behaviour users are asked to monitor and give feedback on. Users give feedback on whether the behaviour aligns with a particular value (e.g., picking apples but being "respectful" to neighbours).

The input to our system is a value that the user wants the agent to adopt. For example, a user could specify to the system "I would like the agent to act respectfully" through a chat interface, as shown in Figure 3. Given this value specification, we collect a large pool

Interactive-Reflective Dialogue Alignment

Figure 2: Overview of the RLHF pipeline with *Interactive-Reflective Dialogue Alignment.*

Figure 3: System screenshot showing the user-facing chat interface of the *Interactive-Reflective Dialogue Alignment* system.

of trajectories (sequences of actions the agent took in the environment). This pool can be collected by taking a portion of an existing dataset or by running random rollouts in a simulator. The system then selectively samples a small number of trajectories where the agent behaves in diverse ways to present to the user for feedback (**trajectory assessment**). Based on this feedback, the system hypothesizes about the rationale behind the user's assessments (e.g., the user thinks the agent's behaviour is disrespectful because it wandered into another agent's yard). This hypothesis is presented to the user along with alternative perspectives the user could consider to prompt them to reflect on their current value definition (**rationale reflection**). Upon reflection, the user can opt to re-explain the trajectories they initially saw with their new perspective. This iterative process is called the preference clarification loop (shown in Figure 1).

All of the information collected from the user through this iterative feedback loop is then used to build an initial language-based

reward model. The initial reward model is refined by selectively picking trajectories that the model is most uncertain about and querying the user for feedback (**uncertainty reduction**). Each time the user explains one of the trajectories the model is uncertain about, this information is added to the reward model, thus making it less uncertain about similar trajectories. This process is repeated until the overall uncertainty is below a certain threshold. In the end, the system generates a language-based reward model that can effectively administer rewards to any agent when it acts in accordance with the user's value definitions (**reward modelling**). Depending on the complexity of the desired behaviour and environment, this reward model can then be used to either train a more efficient reward model or to train an agent directly.
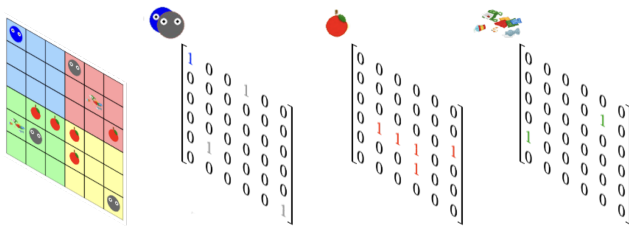
Having described, at a high level, the system and user flow, we will now describe various user flow steps (those in bold) in more detail, namely trajectory assessment, rationale reflection, uncertainty reduction and reward modelling.

3.2.1 Trajectory Assessment. Our system initially lacks information about the user's preferences, and thus, we employ *diversity-based sampling* [52] to find a set of trajectories where the agent exhibits a diverse set of behaviours in varying situations. To do so, we begin by sampling a large pool of over 1000 trajectories. We turn each trajectory into a numerical array that gives a full representation of the state of the environment at each time step. For example, in a 30-step trajectory in the multi-agent apple farming environment, we can encode each step as multiple arrays of the same shape as the grid. In each array, for a step, we can encode the position of a certain type of entity within the environment. The numerical encoding for a single timestep is shown in Figure 4.

With the pool of numerically encoded trajectories in hand, we perform k-means clustering to split the trajectories into k distinct clusters. Since we want to query the user about a diverse set of trajectories in this initial phase, we select one trajectory from each of the k clusters. In particular, for each group, we select the trajectory closest to the arithmetic mean, or the centroid, of all the trajectories in the cluster.

The system implements item-based preference elicitation by asking the user to explain whether the agent is aligned with their preferences in each of the k trajectories, one at a time, using 1-3 sentences. For instance, if a user were trying to train an agent to
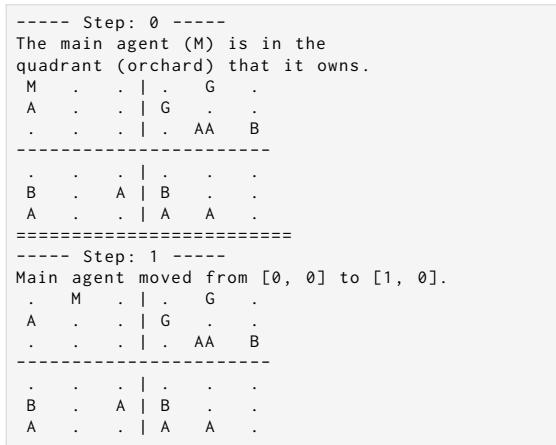
Figure 4: A numerical encoding of one timestep of the multi-agent apple farming environment. Agent positions are encoded in one array, apple positions in a second array, and garbage positions in a third.

behave respectfully, the system would ask, "Does the agent act respectfully? Please explain your reasoning using 1-3 sentences."

3.2.2 Rationale Reflection. During the trajectory assessment process, the system intermittently interjects a dialogue to engage the users to reflect on their value definition. To do this, we use a large language model (GPT-4 Turbo) to make a hypothesis about what features the user is basing their decisions on and other features the user could consider. To do so, we parse each trajectory into an ASCII representation (see Figure 5 for an example) and pass this along with the user's explanation to the LLM. In the prompt, we ask the LLM to make a hypothesis about what features the user is basing their decisions on and to offer alternative features the user could consider.

```
----- Step: 0 -----
The main agent (M) is in the
quadrant (orchard) that it owns.
 M   .   . | .   G   .
 A   .   . | G   .   .
 .   .   . | .   AA   B
-----------------------
 .   .   . | .   .   .
 B   .   A | B   .   .
 A   .   . | A   A   .
=========================
----- Step: 1 -----
Main agent moved from [0, 0] to [1, 0].
 .   M   . | .   G   .
 A   .   . | G   .   .
 .   .   . | .   AA   B
-----------------------
 .   .   . | .   .   .
 B   .   A | B   .   .
 A   .   . | A   A   .
```

Figure 5: ASCII encoding of two timesteps of a trajectory of the multi-agent apple farming environment.

For example, in the multi-agent apple farming environment, where an agent owns one of four orchards in the environment, the user may base their definition of "respectful" behaviour on whether the agent stays in its own orchard. A simplified example of a response from the system could be:

```
Based on your explanations, it seems as
though a key factor in determining whether
the agent's behaviour is respectful or not
is whether the agent stays in its own or-
chard. You could also consider:
(1) Whether the agent helps pick up garbage
(2) Whether the agent steals apples from other
    agents
```

The user is then prompted to explain if the hypothesis the system made is correct and if the other features should be considered. This feature-based preference elicitation step is important for a number of reasons. First, understanding the features on which the user bases their decisions allows for improved generalization of the reward model. The system can use these features, or patterns of behaviour, as decision criteria when assessing whether the agent's behaviour in a trajectory aligns with the user's value definition. Second, this step encourages the user to reflect on their definition and consider alternative perspectives. This reflection can assist users in gaining a clearer understanding of the specific behaviour they want the agent to demonstrate. Third, if the diversity-based sampling in the initial item-based preference elicitation phase does not select any trajectories where a certain behavioural feature is present that would be important to the user, the alternative or additional features proposed by the system can help uncover these. This can help the system gain a more holistic understanding of the user's preferences, even when only querying the user about a handful of items.

If the user updates the features they think are important based on the system's proposed alternatives, they can optionally re-assess the initial k trajectories in the preference clarification loop. This allows users to iteratively reflect on and specify what features or behavioural patterns they would like the agent to embody. This refinement process simultaneously helps the user better understand their own preferences and helps the system gain a clearer understanding of the user's true intent.

3.2.3 Uncertainty Reduction. After the user has provided feedback through the alternating processes of trajectory assessment and rationale reflection, we implement another phase of item-based preference elicitation to create and refine the reward model. This approach focuses on querying the user about specific items (trajectories) to improve the model's understanding of user preferences.

To begin with, we use all of the information collected in the trajectory assessment and rationale reflection phases to create an initial reward model. The reward model can be thought of as a classifier - the input is an ASCII representation of a trajectory, and the output should be a 1 (reward) if the agent's behaviour aligns with the user's expressed intent and 0 (no reward) otherwise. We use this initial model to assign rewards to a held out pool of trajectories. Our reward model, based on an autoregressive transformer (LLM), allows us to identify trajectories where the model has high uncertainty about the appropriate reward. We do this by comparing token probabilities for positive (reward) and negative (no reward) classifications. For example, if the user is training the agent to be respectful, we compare the probability of the reward model outputting the "respectful" token to the probability of the model outputting the "disrespectful" token. If the probabilities are

close to one another, then the model is less certain, and if they are very different (e.g. 0.99 and 0.01), then the model has high certainty about whether the agent acts according to the user's intent. The "confidence" of the model can be thought of as the absolute value of the difference between the two token probabilities.

The item-based preference elicitation process for refining the reward model proceeds as follows: (1) Identify the trajectory in the pool where the model has the lowest confidence. (2) Query the user to explain this trajectory. (3) Add the ASCII representation of the trajectory and the user's explanation to the reward model. (4) Repeat steps 1-3 until the model's confidence for each trajectory in the uncertainty pool exceeds a minimum confidence threshold, ϵ.

This iterative process reduces the model's uncertainty about the user's value definition by focusing on specific items (trajectories) where additional user input is most beneficial.

3.2.4 Reward Modelling. Once the feedback from the user has been collected, the last step is to create a reward model that can give feedback to the learning agent on behalf of the human. Creating a reward model that can issue rewards on behalf of humans is important as, depending on the complexity of the environment and the desired behaviour, it can take millions [55] or even billions [3] of timesteps for an agent to learn the desired behaviour. The reward model can act on behalf of the user, thus relieving the user of giving feedback during agent training.

Our in-context, language-based reward model functions as a classifier, evaluating agent behaviour based on user-defined values. The model receives as input a trajectory encoded in ASCII format, a process detailed in Appendix B. This encoding preserves essential spatial and temporal information while enabling efficient processing by language models (see Figure 5 for an example). The model leverages two key elements to classify a trajectory: (1) The encoded trajectory and (2) the user feedback collected during dialogue-based preference elicitation.

Using this information, we prompt the LLM to assess whether the agent's behaviour in the given encoded trajectory aligns with the user's expressed intent. The LLM's output is then parsed into a binary classification: 1 if the behaviour aligns with user intent, 0 otherwise.

To achieve this, we create a prompt that includes

(1) a description of the environment and the ASCII characters used in the encoding,
(2) the information gained from the user during the dialogue-based preference elicitation,
(3) a description informing the language model that its goal is to predict whether the agent's behaviour is aligned with the user's value definition,
(4) an ASCII representation of the trajectory that the model is to label,
(5) text to encourage the model to engage in chain-of-thought reasoning,
(6) and instructions about how the LLM is to format its response so that the output can be programmatically parsed.

Two major findings about LLM capabilities inspired the design of this prompt. First, LLMs are effective in-context, few-shot learners [14]. This means that a trained LLM, with fixed parameters, can learn new patterns by including the beginning of a pattern as

context and asking the model to continue the pattern. In settings similar to ours, this has been shown to be far more sample efficient than traditional supervised learning [37]. The pattern of interest, in our case, is what behaviour the user deems to be aligned with their value definition. The second finding is that prompting LLMs to engage in chain-of-thought reasoning can greatly increase their performance [36]. The idea is that the model forms an argument first and then, due to the auto-regressive nature of language models, uses the argument it made to determine a final answer.

Since our LLM-based reward model receives a full trajectory as input and outputs a reward, it can deal with non-Markovian rewards. This means that the reward model can evaluate the agent's behaviour over multiple time steps when determining the reward. We hypothesized that this would be important for capturing users' preferred agent behaviour, and this was supported by our empirical findings discussed in Section 7.4.1.

4 STUDY DESIGN & METHODOLOGY

The goal of our system, *Interactive-Reflective Dialogue Alignment*, is to learn individual value definitions. We evaluated our system in two studies by comparing to another language-based reward modelling pipeline and to supervised learning. Study 1 investigates the utility of our system for learning about participants' definition of *respectful* agent behaviour. 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 various language-based reward models, supervised learning baselines, and for a test set to evaluate these methods.
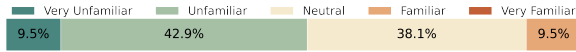
4.1 Environments

In Study 1, the multi-agent apple farming grid world, which was described in the beginning of subsection 3.2, was used. Participants were asked to evaluate if the agent was acting respectfully.

In Study 2, we used the Moral Machine environment [5]. The Moral Machine dataset is based on a simulated environment that presents ethical dilemmas faced by autonomous vehicles. In each scenario, a self-driving car encounters an unavoidable accident and must choose between two outcomes: staying on course or swerving. Each decision results in different consequences for the individuals involved. The environment populates these scenarios with characters chosen from 20 predefined types, including pedestrians, passengers, and various demographic groups such as children, adults, elderly, and animals. Each outcome features between one and five characters. The scenarios explore nine key ethical dimensions: the nature of the vehicle's intervention, the relationship of individuals to the vehicle (pedestrians or passengers), the legality of pedestrians' actions, and various attributes of potential victims, including age, gender, social status, physical fitness, number, and species (human or pet). Participants were asked to choose whether the vehicle should stay or swerve in each scenario presented to them.

4.2 Participants

In Study 1, we recruited 21 participants from our institution (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. The Likert-scale data is visualized in figure 6 and highlights that more than half of participants were "unfamiliar" or "very unfamiliar" with reinforcement learning.

Figure 6: Participant familiarity with reinforcement learning in Study 1.

In Study 2, we recruited 9 participants from our institution (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" (1) to "very familiar" (5), the mean level of familiarity was 3.55, with the mode and median being 3. The Likert-scale data is visualized in figure 7.

Figure 7: Participant familiarity with reinforcement learning in Study 2.

4.3 Procedure

Participants used our interface to specify how they would like the agent to act via our dialogue-based preference elicitation process. Participants then labelled 50 scenarios and were interviewed.

Introduction (∼5min) - After completing the consent form and the demographic questions, the mechanics of the environment were fully explained to the user. We explained the environment mechanics as thoroughly as possible so that differences observed between participants stemmed from their opinions, not hidden assumptions about the environment.

Dialogue - The participant began by conversing with the system about the agent's behaviour following the dialogue structure presented in Section 3.2. To control the amount of time users spent on the activity, we limited the user to one preference clarification loop and one uncertainty reduction loop.

Labelling - Following the participants' dialogue interaction with the system, participants labelled 50 scenarios. Each participant labelled the same scenarios, which allowed us to assess how much the participants agreed on the labels.

Semi-structured Interview (∼10min) - After completing the labelling task, each participant was asked whether they felt they were able to give the system a good understanding of their decision-making if they were always able to articulate why they chose a label, if it was ever difficult to decide, and if they thought their labelling behaviour changed over time. The interview aimed to determine two things. First, it sought to determine if participants could verbally explain their definition of aligned behaviour. Second, it attempted to discover whether these definitions changed over time and what factors caused any changes.

4.4 Baseline Comparisons

We compared our system, *Interactive-Reflective Dialogue Alignment*, to several baselines to evaluate its effectiveness. These baselines include another language-based system and various supervised learning approaches.

4.4.1 Language-Based Baseline (L^B). Kwon et al. [37] proposed a reward modelling 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 slightly 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-based sampling procedure described in Stage 1 of Section 3.2.1.

Since the baseline pipeline L^B is a subprocess of our full system, *Interactive-Reflective Dialogue Alignment*, the user only interacts with our system. However, when forming the reward model for L^B, we only include the information collected from the user during the trajectory assessment phase (described in Section 3.2.1) before they have done rationale reflection, mirroring the pipeline proposed in [37].

4.4.2 Supervised Learning Baselines. We compared our approach to supervised learning methods using neural networks. These baselines include both individual models trained separately for each participant and collective models trained on aggregated data from all participants. For more details on the architecture and training of these models, see Appendix C.

In Study 1, we employed multi-layer perceptron (MLP) models: individual models (MLP_i^{ind}) for each participant i, and a collective model (MLP^{col}) using aggregated data from all participants. These models used input based on the grid map encoding (see Figure 4).

For Study 2, we compared to both MLP and convolutional neural network (CNN) models. We used the same MLP architecture as in Study 1, but with a 26-dimension vector input representing Moral Machine scenarios described in Appendix B. We also introduced CNN models, both individual (CNN_i^{ind}) and collective (CNN^{col}), which use image representations of scenarios as input.

Across all supervised learning models, we utilized 30 scenarios per participant for training, selected from the 50 scenarios annotated during the labelling phase. By comparing *Interactive-Reflective Dialogue Alignment* to these baselines (L^B, MLP_i^{ind}, MLP^{col}, CNN_i^{ind}, and CNN^{col}), we aim to evaluate the efficacy of our language-based reward modeling approach and the value added by our reflective dialogue process.

4.5 Analysis

Our analysis aims to answer the following three questions:

RQ1: Do value definitions significantly vary between participants?
RQ2: Does structured reflection enhance language-based reward modelling?

RQ3: When is individualized language-based reward modelling effective?

To address these 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. The following subsections detail our analytical methods, each designed to provide insights into one or more of our research questions.

4.5.1 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 [39]. Generally, kappa statistics below 0 indicate "poor" agreement and kappa statistics above 0.8 indicate "nearly perfect" agreement [39]. This analysis directly addresses **RQ1** by quantifying how much participants agree when labelling examples. Low agreement suggests diverse value definitions, while high agreement indicates more uniform values across participants. It also informs **RQ3** by indicating when individualized approaches might be more beneficial than collective ones.

4.5.2 Evaluation of Language-Based Reward Model Performance. Our evaluation compares our system's performance against the baseline language-based reward modelling system that does not leverage dialogic (level 3) reflection. We use a performance metric, P, calculated for each participant for both systems. We denote P_i^{IRDA} and P_i^B as the performance metrics for participant i on our system and the baseline system, respectively. The choice of performance metric varied between studies: Study 1 used balanced accuracy due to high class imbalance, while Study 2 used accuracy.

For each participant, we generate rewards (labels) using both systems for the 20 scenarios participants labelled that were not used for training the SL baselines. We then calculate P for each system per participant. This process yields n pairs of P values, where n is the total number of participants. 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 $\Delta P_i = P_i^{\mathrm{IRDA}} - P_i^B$ and bootstrapped these differences in the same way.
(3) We compared the P values for each system using the Wilcoxon signed-rank test. This non-parametric test was chosen over parametric alternatives as it is less prone to false positives and more robust when the data distribution is unknown or skewed [13].

This analysis primarily addresses **RQ2** by comparing the performance of systems with and without dialogic reflection, allowing us to assess if structured reflection enhances language-based reward modelling.

4.5.3 Comparison to Supervised Learning. We compared our language-based systems to traditional supervised learning approaches. Both the individual models ($\mathrm{MLP}_i^{\mathrm{ind}}$ and $\mathrm{CNN}_i^{\mathrm{ind}}$) and the collective models ($\mathrm{MLP}^{\mathrm{col}}$ and $\mathrm{CNN}^{\mathrm{col}}$) 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, representing the performance of the individual and collective models, respectively. To ensure robust statistical analysis, we bootstrapped these values with replacement using 10,000 resamples.

This comparison helps answer **RQ3** by revealing when language-based methods outperform supervised learning approaches under various conditions. It also informs **RQ1** by showing if individualized language-based models consistently capture personal value definitions better than collective supervised models.

4.5.4 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.

This analysis directly addresses **RQ1** by identifying different features and criteria used by participants, providing rich evidence of how value definitions differ.

4.5.5 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 [33]. We computed the Jaccard similarity coefficient for every possible 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. To ensure robustness, we used bootstrapping with 10,000 resamples (sampling with replacement) to determine the 95% confidence interval for this mean Jaccard similarity coefficient.

This quantitative measure helps answer **RQ1** by providing a numerical representation of how much participants' decision-making features overlap. Low similarity scores strongly support the notion that value definitions vary significantly between participants. It also informs **RQ3** by providing a numerical indicator for when individualized vs. collective approaches might be more appropriate.

4.5.6 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 [12]. 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.

This analysis addresses **RQ2** by revealing if reflection enhanced participants' ability to articulate their values, which is crucial for effective language-based reward modelling. It also informs **RQ3**

Table 1: Mapping of Analysis Methods to Research Questions

Analysis Method (§)	RQ1	RQ2	RQ3
Inter-Annotator Agreement (4.5.1)	✓		✓
Evaluation of Language-Based Reward Model Performance (4.5.2)		✓	
Comparison to Supervised Learning (4.5.3)	✓		✓
Qualitative Analysis of Participant Decision Making (4.5.4)	✓		
Analysis of Feature Similarity Between Participants (4.5.5)	✓		✓
Thematic Analysis of Interview Data (4.5.6)		✓	✓

as language-based reward modelling requires that participants can articulate their value definition to be effective.

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 (33.3̄3̄%) entered the *preference clarification loop* for one iteration.

5.0.1 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 [39]. The Fleiss' kappa statistic of 0.336 we observed lends credence to the idea that human values and preferences are subjective and personal.

5.0.2 S1 – Evaluation of Language-Based Reward Model Performance. On average, the reward models produced by our pipeline (IRDA) received significantly higher balanced accuracy 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. The distributions of the balanced accuracies for each pipeline are visualized in the left frame of Figure 8, and the distribution of the per participant difference in balanced accuracy is shown in the right frame.

5.0.3 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 9 illustrates the relationship between model performance and the number of samples provided per participant.
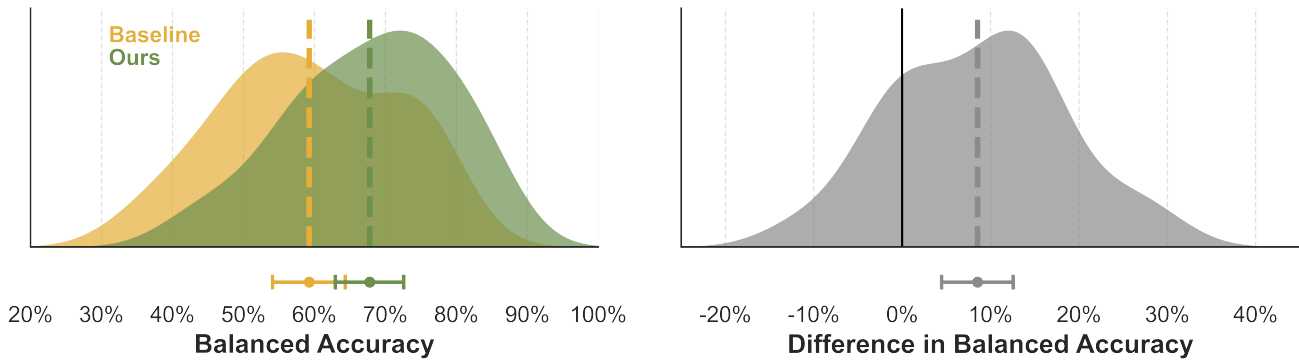
5.0.4 S1 – Qualitative Analysis of Participant Decision Making. While our system is designed to align AI agents with a range of user-defined values, our evaluation concentrated on the singular value of respect. Focusing on a single value limited the variability in the study and allowed us to assess the personal and subjective nature of even a single value.

We analyzed the users' conversations with the system in Study 1 to understand what each participant thought respectful behaviour in this environment was. We found 12 behavioural features that
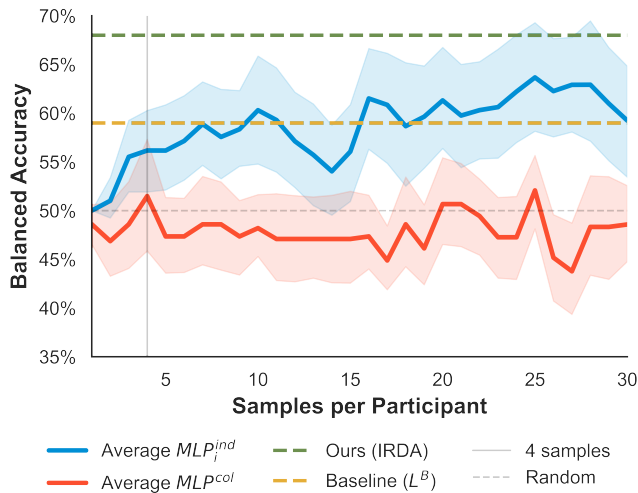
participants used to explain whether they thought the agent was respectful. Interestingly, P1 only used one behavioural feature, namely whether the agent stayed in the quadrant it owned, to determine whether the agent acted respectfully. In contrast, P5, P7, and P10 explained whether the agent acted respectfully using seven features. Overall, only one pair of participants used the same features to make their decisions. Further, participants combined the features in diverse ways, such as forming hierarchies of importance or combining them conditionally. The 12 features we identified are:

(1) **Stays in Own Quadrant:** Whether the main agent stays in the orchard (quadrant) they own.
(2) **Interferes With Others:** This feature evaluates a relatively wide range of behaviours, including if the agent appeared to attempt to block another agent, following another agent, or getting too close to another agent.
(3) **Task Completion:** Whether the agent works toward or completes the task the participant thought the agent should be doing. For example, an agent picking up all apples or garbage in their quadrant.
(4) **Picks Up Own Garbage:** Whether the agent picks up garbage in their own orchard (quadrant)
(5) **Picks Up Others' Garbage:** Whether the agent picks up garbage in other agents orchards
(6) **Tit for Tat Behaviour:** Some participants explained that a given behaviour was respectful if another agent had done it to them first, such as entering their quadrant.
(7) **Taking Others' Apples**: Whether the agent picks up apples in orchards it does not own.
(8) **Eats Own Apples:** Whether the agent eats any apples in their own quadrant.
(9) **Picks Up Garbage Before Apple:** Whether the agent picks up garbage before eating apples in their own or other quadrants.
(10) **Efficiency:** Whether the agent moves around without collecting apples or garbage or makes repetitive, futile movements
(11) **Time in Others' Quadrants:** The duration of time spent in other agents' orchards.
(12) **In Quadrant While Owner was Gone:** If the agent entered another agent's quadrant while they were gone.

Among the features, the most commonly used was **Stays in Own Quadrant**, but participants often found other features more important. Some features are temporally static, while others span multiple steps. For example, whether an agent is in its quadrant can be determined at a single time, but whether an agent picked up

Figure 8: (Left) Distributions of balanced accuracies for language-based reward models: our pipeline (IRDA) vs. baseline (L^B) in Study 1. (Right) Distribution of per-participant differences in balanced accuracy ($P_i^{\text{IRDA}} - P_i^B$) between IRDA and baseline models in Study 1.

Figure 9: Balanced accuracy of different models as a function of samples per participant in Study 1. The blue line represents the average individual MLP model (MLP$^{\text{ind}}$), while the red line shows the collective MLP model (MLP$^{\text{col}}$). Our proposed system (IRDA, green dashed line) and the baseline (L^B, yellow dashed line) were trained on 4 samples per participant, as indicated by the vertical gray line. The collective model was trained on 21 times the number of samples shown on the x-axis due to the 21 participants in Study 1. Shaded areas represent 95% confidence intervals. The gray dashed line indicates random performance.

garbage before eating an apple must be determined over multiple steps. The features each participant used are shown in Figure 10.

5.0.5 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 $(0.333, 0.3813)$.

5.0.6 S1 – Thematic Analysis of Interview Data. Through our thematic analysis, we found two main themes: (1) participants' definitions of respect evolved throughout the activity, and (2) The system's hypothesis and alternative perspectives had a specific impact on this evolution.

Evolving Definitions of Respect. The interview data illustrates that participants' understandings of respect evolved significantly through interaction with the system, particularly influenced by exposure to examples and the system's alternative features. This finding aligns with previous work in consumer research, which found that engaging users in reflection and allowing them to make realistic decisions increases the accuracy of their reported preferences [31]. Initially, many participants held simplistic definitions of respect, often related to spatial boundaries or specific tasks like collecting one's own apples (P4, P6, P7, P10, P19). However, these definitions became more nuanced as participants engaged with the system. The evolution was often spurred by witnessing examples that challenged their initial views and contemplating alternative definitions presented by the system (P3, P8, P10, P13, P18, P19, P20). This highlights a process where the system's feedback and the act of labelling examples prompted participants to refine and sometimes expand their conceptualization of respect, moving beyond their initial assumptions.

Specific Impact of System Hypothesis. Another recurrent theme was the specific influence of the system's hypotheses and feedback on shaping participants' conceptions of respect. Participants reported that the system's presentation of alternative features and hypotheses prompted reevaluation, clarification and, in some cases, a significant revision of their definitions of respect (e.g., P3, P8, P13, P18, P20). For instance, the system's suggestions helped participants to narrow down their considerations (P13, P18) or to think about respect in ways they had not initially contemplated (P3, P20).
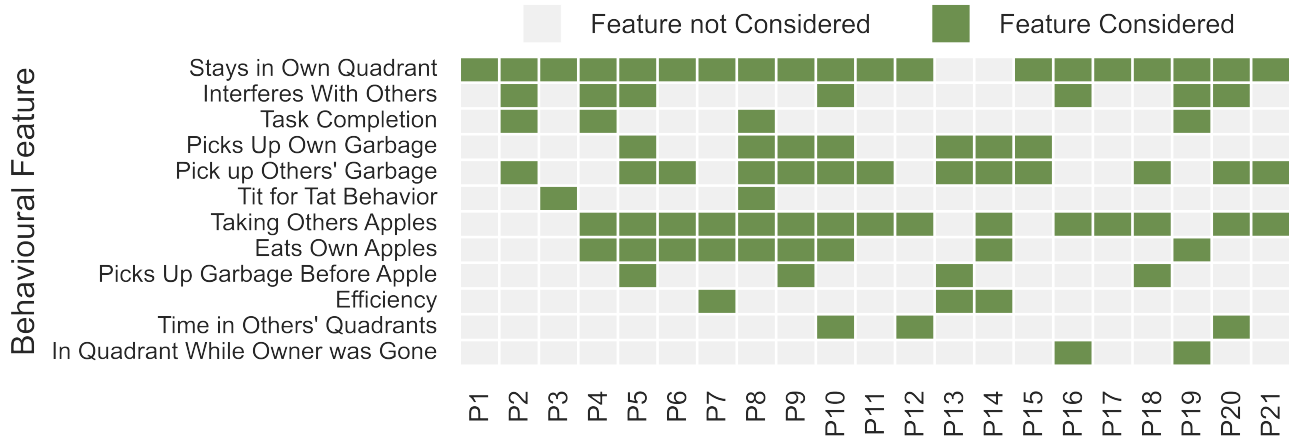
Figure 10: Behavioural features participants used to decide whether the agent was acting respectfully in Study 1.

While the system's alternative features did not always change participants' minds (e.g., P19), they played a role in the iterative process of refining participants' understandings, illustrating the value of engaging the user in 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 clarification loop* for one iteration.

6.0.1 S2 - Inter-Annotator Agreement. We observed a Fleiss' kappa value between all participants' labels on the 50 labelled trajectories of $\kappa = 0.460$, indicating "moderate" (higher than "fair") agreement among participants [39].
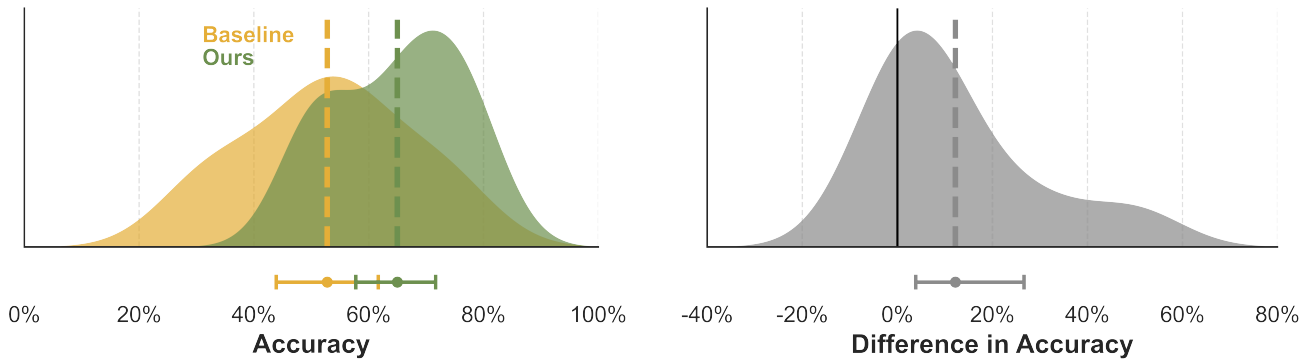
6.0.2 S2 – Evaluation of Language-Based Reward Model Performance. On average, the reward models produced by our pipeline (IRDA) received significantly higher accuracy scores (measured in percentages) than the baseline system (L^B) by 12% (95% CI: [4%, 27%], M = 65% vs. M = 53%, p=.05). This adds more evidence in favour of the effectiveness of structured reflection. The distributions of the balanced accuracies for each pipeline are visualized in the left frame of Figure 11. The distribution of the per-participant difference in balanced accuracy is shown in the right frame.

6.0.3 S2 – Comparison to Supervised Learning. With all 30 training samples, the average accuracy of the individual MLP models (MLP_i^{ind} was 79% (95% CI: [74%, 84%]) while the collective model (MLP^{col}) achieved 77% (95% CI: [75%, 78%]). The left frame Figure 13 illustrates the relationship between model performance and the number of samples provided per participant for the MLP models. For the CNN models, with all 30 training samples, the average accuracy
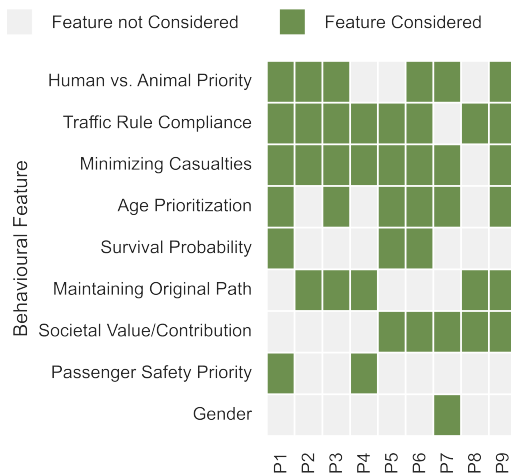
of the individual models (CNN_i^{ind}) was 67% (95% CI: [61%, 73%]) while the collective model (CNN^{col} achieved 77% (95% CI: [70%, 83%]). The right frame of Figure 13 illustrates the relationship between model performance and the number of samples provided per participant for the CNN models. These results point to two things: first, when participant agreement is high, collective methods may outperform individualized methods. Second, when agreement is high and the learning problem becomes more difficult (e.g., the CNN models with image input), pooling of samples is beneficial.

6.0.4 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 features we identified are:

(1) **Human vs. Animal Priority:** Whether human lives are prioritized over animal lives.
(2) **Traffic Rule Compliance:** The consideration of the legality of pedestrians' actions.
(3) **Minimizing Casualties:** The aim to minimize the total number of fatalities.
(4) **Age Prioritization:** The preference for saving younger people over older people.
(5) **Survival Probability:** The consideration of the likelihood of survival for different individuals based on factors like physical fitness.
(6) **Maintaining Original Path:** The preference for the vehicle to stay on its original course rather than swerving.
(7) **Societal Value:** The consideration of perceived societal value or potential future contributions of individuals.
(8) **Passenger Safety Priority:** The prioritization of the safety of the vehicle's passengers over pedestrians or other road users.
(9) **Gender:** The consideration of the gender of potential victims in the decision-making process.

Figure 11: (Left) Distributions of accuracies for language-based reward models: our pipeline (IRDA) vs. baseline in Study 2. (Right) Distribution of per-participant differences in accuracy ($P_i^{\text{IRDA}} - P_i^B$) between IRDA and baseline models in Study 2.

Figure 12: Behavioural features participants used to decide what the autonomous vehicle should do in Study 2.

Like Study 1, participants combined and used these features in various ways.

6.0.5 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 $(0.403, 0.526)$.

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

Decision-making Evolution. Our results reveal a divergence in how participants' decision-making processes evolved throughout the study. Some participants reported that their approach changed as they encountered a wider range of scenarios. For instance, P3

noted that "as more cases came up, I realized the need to consider new factors when the initial factors were equal between groups." This reiterates our finding from Study 1 that exposure to diverse situations can prompt users to refine and expand their decision-making criteria. Conversely, other participants maintained consistent rules throughout the study. P2 stated that their "rules remained consistent throughout," indicating that some users may approach such tasks with pre-established ethical frameworks that remain stable across various scenarios.

Intuition vs. Explicit Reasoning. Our findings reveal an interplay between explicit reasoning and intuition. Most participants felt capable of articulating their decision-making process. Still, the emergence of intuition-based decisions in particularly challenging scenarios, as reported by P6 and P7, underscores the complexity of ethical reasoning. P7 mentioned relying on "first instinct" or "vibes" for 3 or 4 especially difficult scenarios.
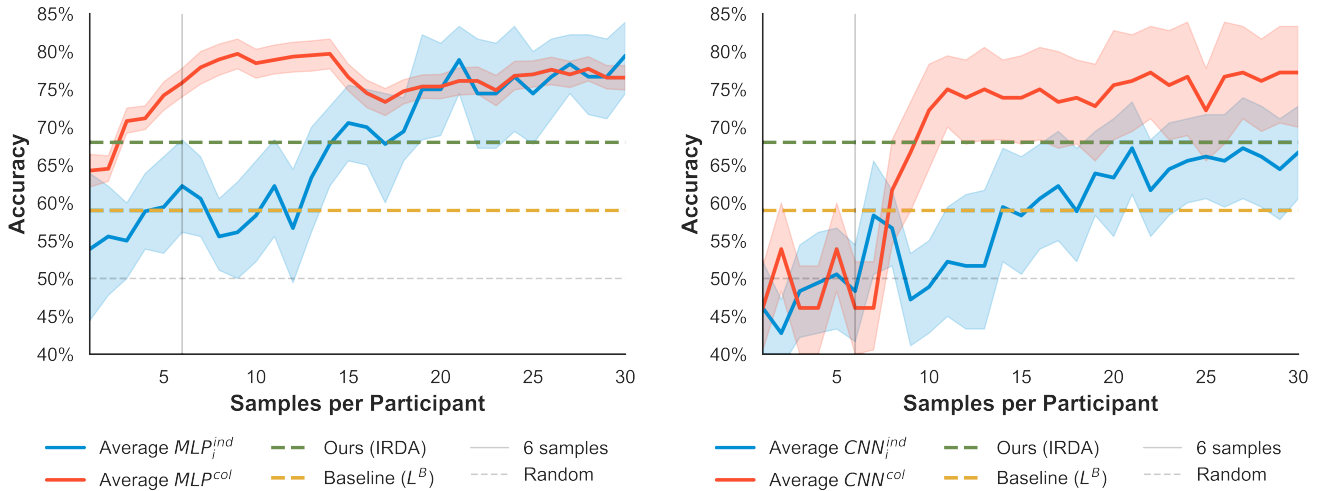
7 DISCUSSION

Our evaluation of *Interactive-Reflective Dialogue Alignment* yields insights that challenge fundamental assumptions about human values and AI alignment. These findings have implications for the design of AI systems that aim to respect and represent diverse values and preferences.

7.1 The Diversity of Human Values (RQ1)

In both Study 1 and 2, we observed heterogeneity in participants' value definitions. However, participants' value definitions varied to a greater degree in Study 1.

The diversity in participants' interpretations of respectful behaviour, observed in Study 1, suggests that values like respect are not universal constants but rather individualized constructs. This heterogeneity manifested in several ways: The identification of 12 distinct behavioural features used by participants to evaluate respectful agent behaviour reveals the multifaceted nature of respect. Further, the unique combinations of these features used by participants highlight the individualized nature of value interpretation. This finding extends previous work on value pluralism in moral

Figure 13: Comparison of model accuracies as a function of samples per participant in Study 2. (Left) Performance of MLP-based models: average individual MLP (MLP$^{\text{ind}}$, blue), collective MLP (MLP$^{\text{col}}$, red), our IRDA approach (green dashed), and baseline L^B (yellow dashed). (Right) Performance of CNN-based models: average individual CNN (CNN$^{\text{ind}}$, blue), collective CNN (CNN$^{\text{col}}$, red), IRDA, and baseline. Both panels show confidence intervals (shaded areas), the 6-sample training point for IRDA and baseline (vertical gray line), and the random performance level (gray dashed). IRDA consistently outperforms other approaches across model architectures.

philosophy [10], suggesting that even seemingly universal values are subject to significant individual variation.

The "fair" inter-annotator agreement ($\kappa = 0.23$) and relatively low feature usage similarity ($J = 0.357$) in Study 1 further underscores this diversity. This level of disagreement, observed when participants evaluated identical agent behaviours, implies that discussions about values may often involve fundamentally different conceptualizations masked by shared terminology.

Perhaps the most compelling evidence for this diversity in Study 1 comes from comparing the performance of individual models (MLP$^{\text{ind}}_i$) with the collective reward model (MLP$^{\text{col}}$). Individual models achieved an average balanced accuracy of 59%, demonstrating that above-random performance is possible in this task with only 30 samples. However, despite access to 630 samples, the collective reward model could not surpass random performance. This contrast suggests that the diversity in value interpretations is not simply noise that can be averaged out with more data. Instead, it indicates genuine and significant differences in how individuals interpret values.

Study 2, in comparison, revealed more homogeneous opinions, with higher inter-annotator agreement ($\kappa = 0.460$) and greater feature usage similarity ($J = 0.464$). In accordance with this, the collective models (MLP$^{\text{col}}$ and CNN$^{\text{col}}$) were generally more effective in capturing participants' value definitions than individualized models.

The contrast between Studies 1 and 2, where Study 2 revealed more homogeneous opinions, highlights the context-dependent nature of value diversity. This finding suggests that the degree of consensus about values may vary significantly across different scenarios or domains. These observations challenge the assumption that a universal set of values can be embedded in AI agents in all contexts. Given the diversity in value definitions, personalization or, in multi-stakeholder settings, making informed compromises that respect all stakeholders are both viable alignment strategies. Individualized reward models facilitate both personalization and informed and interpretable compromises.

7.2 The Power of Structured Reflection (RQ2)

The increased performance of our reflection-based system compared to the non-reflective language-based baseline in Studies 1 and 2 demonstrates the effectiveness of structured reflection in preference elicitation and communication. This finding suggests that guiding users to consider alternative perspectives and reflect on their preferences enables a more accurate capture of individual value definitions.

This observation aligns with dual-process theories in cognitive psychology [21], indicating that our reflection process engages System 2 thinking - the deliberate and analytical mode of thought. By prompting users to articulate and justify their value judgments, we facilitate a process of value discovery and refinement.

Qualitative feedback from Study 1 participants provides insight into this process. Users reported that engaging with the system's hypotheses and alternative features allowed them to refine or clarify their understanding of respectful behaviour. This suggests that

the reflection process serves as a form of cognitive scaffolding, supporting users in exploring and articulating their values. Drawing upon research on "designing for reflection" [25], our dialogue system creates opportunities for users to externalize and examine their internal value frameworks.

Intriguingly, Study 2 revealed that our system outperformed the baseline even when participants had more fixed opinions and engaged less with the preference clarification loop. This finding highlights a crucial aspect of our approach: structured reflection enhances preference communication, improving the system's ability to capture nuanced preferences even when users' fundamental views remain unchanged. This aligns with recent work that suggests reflective System 2 thinking can help individuals justify, rationalize, and explain their intuitions [16].

These results suggest that AI systems may need to do more than observe behaviour or collect binary feedback to understand and align with human values. Instead, they may need to engage humans in a process of structured reflection, creating a dialogue that helps both the human and the AI system develop a clearer understanding of the human's values and preferences. Our findings contribute to the ongoing discourse on value alignment in AI, suggesting that effective preference elicitation requires not just sophisticated AI models, but also carefully designed interaction paradigms that support human cognitive processes.

7.3 The Contextual Efficacy of Individualized Language-Based Reward Modeling (RQ3)

Our investigation into the efficacy of individualized language-based reward modelling (RQ3) reveals a nuanced landscape of strengths and limitations. The approach demonstrates remarkable sample efficiency, outperforming individual models trained on 30 samples and collective models trained with 630 samples with just four samples in Study 1. This efficiency stems from our method's structured reflection process and the few-shot learning capabilities of LLMs. However, this efficiency is context-dependent, with effectiveness varying based on preference heterogeneity within the population. Study 1, characterized by higher preference diversity (Fleiss' kappa = 0.336, Jaccard similarity = 0.357), showcased the strengths of our individualized approach. In contrast, Study 2 exhibited more homogeneous preferences (Fleiss' kappa = 0.460, Jaccard similarity = 0.464), revealing conditions where collective supervised learning performed better.

Input representation emerged as an important factor in preference learning, as demonstrated by the performance difference in Study 2 between MLP (using feature-engineered inputs) and CNN (using image inputs) models. This aligns with work on representation learning [8] and underscores the importance of explicit feature identification, a core aspect of our dialogue-based approach, in capturing decision-making processes.

Our findings indicate that individualized language-based reward modelling is most effective under the following conditions: high preference heterogeneity, limited samples per individual, complex learning problems, and scenarios requiring feature discovery. These conditions stand in contrast to those favouring traditional reward modelling methods, which thrive with homogeneous preferences, large datasets, and simpler learning problems.

7.4 Other Findings

7.4.1 Beyond Markov: Capturing Temporal Dynamics in Value Judgments. An important finding from our study was the prevalence of non-Markovian features in participants' evaluations of respectful behaviour. This observation challenges a fundamental assumption in many reinforcement learning systems: that the current state contains all the necessary information to make a decision. Participants often based their judgments on sequences of actions rather than single states. For instance, considering whether an agent picked up garbage before collecting an apple or whether it entered a quadrant previously visited by another agent. These non-Markovian features require information from multiple time steps, highlighting the temporal nature of many value judgments.

This finding suggests that to capture human values accurately, reward models need to consider temporal sequences and historical context, not just instantaneous states. Our approach, which takes entire trajectories as input, naturally accommodates these non-Markovian features, allowing for a more comprehensive understanding of human values.

7.5 Limitations

7.5.1 Implementation Dependence. Some aspects of our system are artifacts resulting from the current capabilities of large language and vision language models. For example, we textually encode trajectories so that they can be input into a language model. We chose this because vision language models performed poorly in the environments we tested. However, this may not be necessary in the future as the capabilities of LLMs and VLMs increase. That said, the core of the system, namely the overall pipeline, is not model dependent, and as capabilities increase, we expect the performance of the pipeline only to increase.

7.5.2 Reward Model vs. Agent Accuracy. We focused our system and evaluation on the feedback collection and reward modelling phases of the RLHF pipeline and did not consider policy optimization (agent training). This approach aligns with recent trends in the field, where reward models themselves are the subject of evaluation [38, 68]. We chose this focus because our core innovation is not related to policy optimization. While this leaves questions open about the behaviour that would result from training an agent with a reward model produced by our system, Kwon et al. [37] found that RL agent accuracy mimics the reward model accuracy, suggesting that results about reward models are transferable to trained agents.

7.6 Future Work

Our research into *Interactive-Reflective Dialogue Alignment* opens up several promising avenues for future investigation.

7.6.1 Interactive RLHF. In our studies, we looked specifically into reward modelling and did not train agents with the reward models generated by our system. However, each type of reward model, including the ones generated by our system and the ones we compared to, had imperfections. These imperfections can result in imperfect agents. As such, future work can examine how to interleave reward modelling and agent training. In this way, the user can iteratively give feedback directly addressing the trained agent's imperfections.

7.6.2 Hybrid Approaches. Our current work treated language-based reward modelling and supervised learning as distinct approaches. However, there's potential for developing hybrid methodologies that leverage the strengths of both. For instance, our language-based reward modelling pipeline could be used to generate additional training samples for supervised learning, similar to the approach presented by Lee et al. [41].

7.6.3 Aligning to Group vs. Individual Values. While our work focused on capturing individual preferences, many real-world scenarios require alignment with collective or group values. Future research could extend our approach to group alignment by aggregating individual preferences. This direction builds on existing work in collective decision-making and virtual democracy [26, 29, 47, 53], but applies these concepts specifically to the challenge of AI alignment. Key questions in this area include how to fairly aggregate diverse individual preferences in temporally extended scenarios, how to handle conflicts between individual and group values, and how to ensure that the resulting aligned AI systems are acceptable to all stakeholders.

8 CONCLUSION

We developed a system, *Interactive-Reflective Dialogue Alignment*, that aids users with no particular experience in machine learning in designing a reward model that can be used to train agents in alignment with their individual understanding of values. Our findings demonstrate that *Interactive-Reflective Dialogue Alignment* effectively captures diverse and subjective value definitions through an interactive, reflective, language-based process. By enabling users to engage deeply with the nuances of their values, the system provides an individual-conscious approach to AI alignment. The studies on "respect" and ethical decision-making in autonomous vehicles illustrate the system's capability to accommodate a wide range of value-aligned behaviors. This method can be used to personalizes AI agents and as the foundation for interpretable and representative collective alignment strategies

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. 1.

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[3] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. 2022. Melting Pot 2.0. *arXiv preprint arXiv:2211.13746* (2022).

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. https://doi.org/10.48550/arXiv.1606.06565 arXiv:1606.06565 [cs].

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

[6] 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. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. https://doi.org/10.48550/arXiv.2204.05862 arXiv:2204.05862 [cs].

[7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. https://arxiv.org/abs/2212.08073v1

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[9] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. 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 (March 2022), 2:1–2:27. https://doi.org/10.1145/3517233

[10] Isaiah Berlin. 1969. Four essays on liberty.

[11] Erdem Biyik, Fan Yao, Yinlam Chow, Alex Haig, Chih-wei Hsu, Mohammad Ghavamzadeh, and Craig Boutilier. 2023. Preference Elicitation with Soft Attributes in Interactive Recommendation. *arXiv preprint arXiv:2311.02085* (2023).

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

[13] Patrick D Bridge and Shlomo S Sawilowsky. 1999. 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 (1999), 229–235.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs].

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

[16] Dario Cecchini. 2021. Dual-process reflective equilibrium: rethinking the interplay between intuition and reflection in moral reasoning. *Philosophical Explorations* 24, 3 (2021), 295–311.

[17] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 815–824.

[18] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. https://arxiv.org/abs/1706.03741v4

[19] Galileu B Costa Filho, Alexandre S Moura, Paulo R Brandão, Henk G Schmidt, and Silvia Mamede. 2019. Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. *Perspectives on Medical Education* 8 (2019), 230–236.

[20] Daniel Dewey. 2014. Reinforcement learning and the reward engineering principle. In *2014 AAAI Spring Symposium Series*.

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

[22] Tom Everitt and Marcus Hutter. 2016. Avoiding wireheading with value reinforcement learning. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9*. Springer, 12–22.

[23] Rachel Aparecida Ferreira Fernandes, Leandro Fernandes Malloy-Diniz, Marcos Carvalho de Vasconcellos, Paulo Augusto Moreira Camargos, and Cássio Ibiapina. 2021. Adding guidance to deliberate reflection improves medical student's diagnostic accuracy. *Medical Education* 55, 10 (2021), 1161–1171.

[24] Angela Fessl, Oliver Blunk, Michael Prilla, and Viktoria Pammer. 2017. The known universe of reflection guidance: a literature review. *International journal of technology enhanced learning* 9, 2-3 (2017), 103–125.

[25] Rowanne Fleck and Geraldine Fitzpatrick. 2010. 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*. 216–223.

[26] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283 (2020), 103261.

[27] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory* (2013), 55–95.

[28] Maliheh Ghajargar, Mikael Wiberg, and Erik Stolterman. 2018. Designing IoT systems that support reflective thinking: A relational approach. *International Journal of Design* 12, 1 (2018), 21–35.

[29] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[30] Mark P Graus and Martijn C Willemsen. 2015. Improving the user experience during cold start through choice-based preference elicitation. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 273–276.

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

[32] Dirk Ifenthaler. 2012. Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. *Journal of Educational Technology & Society* 15, 1 (2012), 38–52.

[33] Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist* 11, 2 (1912), 37–50.

[34] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. 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 (July 2018), 70:1–70:26. https://doi.org/10.1145/3214273

[35] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. 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 (2018), 1–26.

[36] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[37] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2022. Reward Design with Language Models. In *The Eleventh International Conference on Learning Representations*.

[38] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787* (2024).

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

[40] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. 2009. 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*. 1141–1150.

[41] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267* (2023).

[42] Madelene Lindström, Anna Ståhl, Kristina Höök, Petra Sundström, Jarmo Laaksolathi, Marco Combetto, Alex Taylor, and Roberto Bresin. 2006. Affective diary: designing for bodily expressiveness and self-reflection. In *CHI'06 extended abstracts on Human factors in computing systems*. 1037–1042.

[43] Maria Salamó Llorente and Sergio Escalera Guerrero. 2011. Increasing retrieval quality in conversational recommenders. *IEEE Transactions on Knowledge and Data Engineering* 24, 10 (2011), 1876–1888.

[44] Benedikt Loepp, Tim Hussein, and Jüergen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3085–3094.

[45] Silvia Mamede, Henk G Schmidt, and Júlio César Penaforte. 2008. Effects of reflective practice on the accuracy of medical diagnoses. *Medical education* 42, 5 (2008), 468–475.

[46] Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. 2023. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning*. PMLR, 24555–24591.

[47] Farhad Mohsin, Lei Luo, Wufei Ma, Inwon Kang, Zhibing Zhao, Ao Liu, Rohit Vaish, and Lirong Xia. 2021. Making group decisions from natural language-based preferences. In *Proceedings of the 8th International Workshop on Computational Social Choice (COMSOC)*. 2.

[48] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

[49] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (April 2023), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

[50] Gergely Neu and Csaba Szepesvári. 2009. Training parsers by inverse reinforcement learning. *Machine learning* 77 (2009), 303–337.

[51] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[52] Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*. 79.

[53] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[54] Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. (2009).

[55] Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. 2018. Time limits in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4045–4054.

[56] Sara Price, Yvonne Rogers, Danae Stanton, and Hilary Smith. 2003. A new conceptual framework for CSCL: Supporting diverse forms of reflection through multiple interactions. In *Designing for change in networked learning environments: Proceedings of the International Conference on Computer Support for Collaborative Learning 2003*. Springer, 513–522.

[57] Bilih Priyogi. 2019. Preference Elicitation Strategy for Conversational Recommender System. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, Melbourne VIC Australia, 824–825. https://doi.org/10.1145/3289600.3291604

[58] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Satoshi Nakamura, Milica Gasic, Ingrid Zukerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino (Eds.). Association for Computational Linguistics, Stockholm, Sweden, 353–360. https://doi.org/10.18653/v1/W19-5941

[59] Deepak Ramachandran and Eyal Amir. 2007. Bayesian Inverse Reinforcement Learning.. In *IJCAI*, Vol. 7. 2586–2591.

[60] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)* 54, 9 (2021), 1–40.

[61] Bettina Renner, Michael Prilla, Ulrike Cress, and Joachim Kimmerle. 2016. Effects of prompting in reflective learning tools: Findings from experimental field, lab, and online studies. *Frontiers in psychology* 7 (2016), 820.

[62] Rita Almeida Ribeiro. 1996. Fuzzy multiple attribute decision making: a review and new preference elicitation techniques. *Fuzzy sets and systems* 78, 2 (1996), 155–181.

[63] Yvonne Rogers and Henk Muller. 2006. A framework for designing sensor-based interactions to promote exploration and reflection in play. *International Journal of Human-Computer Studies* 64, 1 (2006), 1–14.

[64] Stuart Russell. 2019. *Human compatible: AI and the problem of control*. Penguin Uk.

[65] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[66] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology/Academic Press* (1992).

[67] Burr Settles. 2009. Active learning literature survey. (2009).

[68] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Understanding Hidden Context in Preference Learning: Consequences for RLHF. In *The Twelfth International Conference on Learning Representations*.

[69] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*. 235–244.

[70] Ibo Van de Poel. 2013. Translating values into design requirements. *Philosophy and engineering: Reflections on practice, principles and process* (2013), 253–266.

[71] Paolo Viappiani, Boi Faltings, and Pearl Pu. 2006. Preference-based search using example-critiquing with suggestions. *Journal of artificial intelligence Research* 27 (2006), 465–503.

[72] Joseph Jay Williams, Tania Lombrozo, Anne Hsu, Bernd Huber, and Juho Kim. 2016. Revising learner misconceptions without feedback: Prompting for reflection on anomalies. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 470–474.

[73] Irmtraud Wolfbauer, Viktoria Pammer-Schindler, Katharina Maitz, and Carolyn P. Rose. 2022. A Script for Conversational Reflection Guidance: A Field Study on Developing Reflection Competence With Apprentices. *IEEE Transactions on Learning Technologies* 15, 5 (Oct. 2022), 554–566. https://doi.org/10.1109/TLT.2022.3207226

[74] Irmtraud Wolfbauer, Viktoria Pammer-Schindler, Katharina Maitz, and Carolyn P Rosé. 2022. A script for conversational reflection guidance: a field study on developing reflection competence with apprentices. *IEEE Transactions on Learning Technologies* 15, 5 (2022), 554–566.

[75] Jiayu Yin, Catherine Gu, Jenny Mar, and Sydney Zhang. 2024. Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. (2024).

[76] Joyce Yukawa. 2003. Co-reflection in online learning environments. *ACM SIGGROUP Bulletin* 24, 3 (2003), 44–49.

[77] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.

[78] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A SYSTEM ALGORITHM

Algorithm 1: *Interactive-Reflective Dialogue Alignment*

Data: Desired behaviour lang(behaviour), number of diversity-based trajectories to query the user about k, confidence threshold ϵ

1   Diversity-based sampling of k trajectories using k-means
2   **for** *i in k* **do**
3   Show video of trajectory i
4   Collect user feedback
5   Add ASCII encoding of trajectory and user feedback to reward model
6   Generate user feature hypothesis and alternative features
7   Collect user reflection
8   Add hypothesis, alternative features, and user reflection to the reward model.
9   **if** *user chooses to enter the preference clarification loop* **then**
10   return to step 2
11   **while** *model confidence* $< \epsilon$ **do**
12   Use uncertainty sampling to select trajectory
13   Show video of trajectory
14   Collect user feedback
15   Add ASCII encoding of trajectory and feedback to reward model
16   Recalculate *model confidence*
17   **return** *Reward model*

B TRAJECTORY ENCODING

B.1 Study 1 - Multi-Agent Apple Farming

During the feedback collection phase, the user converses with the system about how they would like the agent to act. All of the feedback collected from the user is in natural language. Thus, the reward model must be able to make use of natural language information. However, the information about the agent and environment may not be in natural language. For example, in the multi-agent apple farming environment, the information about the environment and the agent's actions is stored in a numerical format that can later be rendered as an RGB array when a human wants to view it. This presents a challenge as the feedback from the user and the information about what the feedback is referring to, namely the agent and the environment, are not in the same format.

One possible solution to this grounding problem is to use a vision-language model (VLM) that can accept both natural language and visual information. However, when we performed preliminary testing, we found that the performance of VLMs was notably poor when attempting to prompt the model to reason about detailed spatial information from reinforcement learning environments such as grid worlds. It is possible that as VLMs improve, they will be able to reason about high-granularity spatial images successfully. Regardless, given current capabilities, we opted to convert the trajectories into ASCII representations, which an LLM can understand.

To do so, we coded each entity in the environment with an ASCII character. For example, for the environment used in our evaluation, the main agent the user was giving feedback on was represented

by an "M", apples, which the agent could collect for reward, by "A", other agents by "B", garbage by "G", empty tiles by ".", and orchard boundaries by "–" and "|". We also annotated each step in the trajectory with the step number and important actions that occurred. Figure 5 shows an example of two timesteps encoded in this way. Notably, all of the information comes from the underlying state of the environment and is programmatically generated. This means a human does not have to convert and annotate each trajectory, which is important for scalability.

B.2 Study 2 - The Moral Machine

B.2.1 Numerical. The Moral Machine dataset is transformed into a 26-dimensional vector representation for each scenario by selecting relevant features from the original dataset described in detail here: https://osf.io/wt6mc. These features include scenario structure elements (Intervention, PedPed, Barrier, CrossingSignal), numeric characteristics (NumberOfCharacters, DiffNumberOFCharacters), and 20 character types (e.g., Man, Woman, Pregnant, etc.). Each moral dilemma compares two outcomes (stay vs. swerve), represented by two rows in the dataset. The vectorization process involves creating comparison pairs based on the 'Intervention' column, then calculating the difference between feature values of the two outcomes. This differencing captures the relative changes in features between outcomes, resulting in a 26-dimensional difference vector. Before use in the Multi-Layer Perceptron classifier, these vectors are standardized using StandardScaler to ensure consistent feature scaling.

```
The brakes of a self-driving car have
failed. The self-driving car can
continue driving straight ahead
or swerve. If the car continues
straight ahead, it will run over
a group of 5 pedestrians who are
crossing against a red do not walk
signal. The group of pedestrians
include:
    - 4 girls
    - A female doctor
If the car swerves, it will run over a
group of 5 pedestrians who are crossing
with a green walk signal. The group of
pedestrians include:
    - 4 boys
    - A male doctor
```
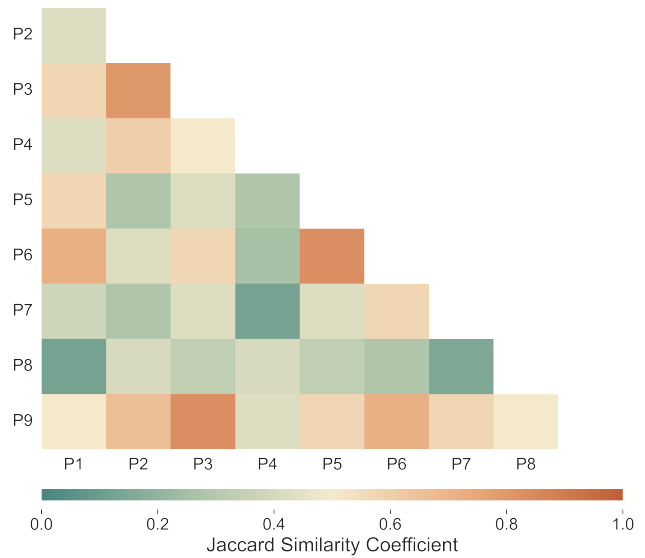
Figure 14: ASCII encoding of a Moral Machine scenario from Study 2.

B.2.2 ASCII. The process of creating natural language descriptions for the Moral Machine scenarios involves converting the raw data into a verbal description. Each scenario is described by presenting the basic dilemma of a self-driving car with failed brakes, followed by the two possible outcomes (continuing straight or swerving). The description includes details about the number and types of characters involved in each outcome, their actions (such as crossing legally or illegally), and any relevant attributes (like profession or age). An example is shown in Figure 14.
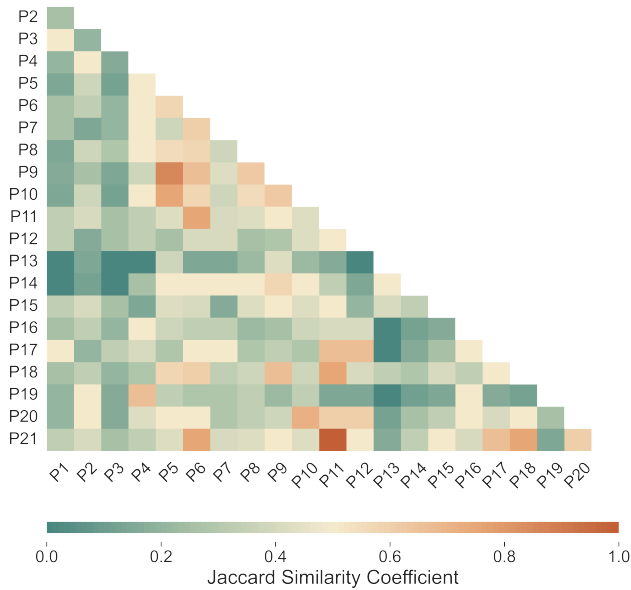
C   DETAILED DESCRIPTION OF SUPERVISED LEARNING BASELINES

The MLP models consist of one hidden layer with 32 neurons, trained with a learning rate of 0.001 and the Adam optimizer. For Study 1, the input was based on the grid map encoding, while for Study 2, a 26-dimension vector representing Moral Machine scenarios was used. The CNN architecture for Study 2 comprises two convolutional layers followed by ReLU activation and max pooling. The first layer has 16 filters, and the second has 32 filters, both with a kernel size of 3 and padding of 1. After flattening, there are two fully connected layers with ReLU activation, reducing the dimensionality to 64 and then to the final output size of 2. These architectures were chosen to balance model complexity with the limited amount of training data available per participant. The use of both individual and collective models allows us to compare personalized preferences with aggregated group preferences across different neural network structures and input representations.

D   JACCARD SIMILARITY VISUALIZATIONS

Figure 16: Heatmap of pairwise Jaccard similarity coefficients between participants (P1-P9) based on their use of decision-making features in Study 2. The Jaccard similarity coefficient quantifies the overlap in features used by each pair of participants, with values ranging from 0 (no overlap) to 1 (complete overlap).

Figure 15: Heatmap of pairwise Jaccard similarity coefficients between participants (P1-P21) based on their use of decision-making features in Study 1. The Jaccard similarity coefficient quantifies the overlap in features used by each pair of participants, with values ranging from 0 (no overlap) to 1 (complete overlap).