Learning the Value Systems of Societies with Preference-based Multi-objective Reinforcement Learning

Andrés Holgado-Sánchez
CETINIA, University Rey Juan Carlos
Madrid, Spain
andres.holgado@urjc.es

Peter Vamplew
Federation University
Ballarat, Australia
p.vamplew@federation.edu.au

Richard Dazeley
Deakin University
Geelong, Australia
richard.dazeley@deakin.edu.au

Sascha Ossowski
CETINIA, University Rey Juan Carlos
Madrid, Spain
sascha.ossowski@urjc.es

Holger Billhardt
CETINIA, University Rey Juan Carlos
Madrid, Spain
holger.billhardt@urjc.es

ABSTRACT

Value-aware AI should recognise human values and adapt to the value systems (value-based preferences) of different users. This requires acquiring computable representations of values, a process that can be prone to misspecification. The social nature of values demands their representation to adhere to multiple users while value systems are diverse, yet exhibit patterns among groups. In sequential decision making, efforts have been made towards personalization for different goals or values from demonstrations of diverse agents. However, these approaches demand manually designed features or lack value-based interpretability and/or adaptability to diverse user preferences.

We propose algorithms for learning models of value alignment and value systems for a society of agents in Markov Decision Processes (MDPs), based on clustering and preference-based multi-objective reinforcement learning (PbMORL). We jointly learn socially-derived value alignment models (groundings) and a set of value systems that concisely represent different groups of users (clusters) in a society. Each cluster consists of a value system representing the value-based preferences of its members and an approximately Pareto-optimal policy that reflects behaviours aligned with this value system. We evaluate our method against a state-of-the-art PbMORL algorithm and baselines on two MDPs with human values.

KEYWORDS

Value Awareness; Value Alignment; Inverse Reinforcement Learning; Multi-objective Reinforcement Learning; Preference-based Reinforcement Learning

ACM Reference Format:
Andrés Holgado-Sánchez, Peter Vamplew, Richard Dazeley, Sascha Ossowski, and Holger Billhardt. 2026. Learning the Value Systems of Societies with Preference-based Multi-objective Reinforcement Learning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/NLVD8864

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/NLVD8864

1 INTRODUCTION

To remain human-aligned, AI agents require awareness of human values such as *benevolence* or *achievement* [29], i.e. the ability to reason with and about the value-alignment of different actions. This allows making explicit the alignment with the *pluralistic value systems* [22] of human agents, manifested as value-based preferences[1].

Value-awareness requires computational representations of values: existing works model human values with, e.g. utility functions [30, 51], reward functions [45, 61], constraints [4, 69] or logic programming [2]. Value systems are typically represented by value orderings [45, 52], weights over value alignment utilities/rewards [22, 25], preferences over alternatives [9, 16] or other methods such as taxonomies [35]. Such models can be applied to choosing value-aligned government policies [25], obtaining aligned behaviour [38, 45] or norm selection in multi-agent systems [1, 30].

However, representing values is challenging, due to their context-sensitive [24], and socially emergent and evolving nature [35, 49]. This often results in misspecification [57]. Also, most of the surveyed works that analyse value-based preference diversity (e.g. [22, 25]) cannot simultaneously represent the various levels of alignment with values demanded by different user cohorts [62].

For Markov Decision Processes (MDPs), reward learning is a promising way to achieve value-alignment [11, 21]. Many works, though, focus on human alignment for the whole society [41], but without awareness of the multiple value-based preferences involved. As such, we argue that value alignment should be treated as a multi-objective problem instead [58]; in particular, encoding separate values as different components of a reward vector allows an AI system to recognise these values and abide with them under varying conditions. Modelling values and value-based preferences of specific users has been approached via multi-objective rewards [45, 61, 66], but these works struggle to instantiate models that concisely represent the diversity of value preferences within human groups.

In this paper, we propose an online preference-based multi-objective reinforcement learning [31] algorithm with clustering that learns the value system of a society in a MDP. This entails learning: a) a social *grounding* (reward vector) that represents the alignment of different behaviours with a given set of values; b) a set of values systems (linear scalarization functions) that represent different clusters of agents with similar value-based preferences; and c)

[1]Here, "value-based preferences", refer to preferences between alternatives based on their alignment [47] with multiple values.

a set of MDP policies, each aligned with one of those value systems. We set out from a *social value system learning* algorithm [16] that learns a) and b) in non-sequential decision-making. Our algorithm asks for pairwise preferences between trajectories (i.e. sequences of observations and actions) in terms of the value systems of each agent (value-based preferences) and in terms of their alignment with the given set of values (value alignment preferences). We evaluate the capabilities of the method to approximate the value systems and behaviours of societies in synthetic MDPs.

The paper is structured as follows. In Section 3 we propose a representation of the value system of a society. In Section 4 we describe the social value system learning problem and propose an algorithm to solve it. In Section 5 we evaluate our methods against baselines and an existing PbMORL algorithm [31]. In Section 6, we provide conclusions, limitations, and suggestions for future work.

2 RELATED WORKS

2.1 Value Awareness, Inference and Learning

Artificial Moral Agents (AMAs) [60] can be classified as explicit or implicit depending on whether their behaviour can be expressed in ethical terms [7]. Implicit systems acquire *(value-)aligned* behaviour through imitation [40, 41] or teaching [14, 28]; explicit ones operationalize values or principles to permit the ethical evaluation of actions [2] or norms [33, 34]. Value-aware systems [29] extend AMAs by explicitly grounding values as desirable goals [50] and respecting agents' *value systems* (preferences over values [46]).

Given the difficulty of eliciting values and value systems, Liscio et al. [23] propose *value inference* as a three-step process, namely: i) *value identification* [24, 40], which finds the values relevant to a society; ii) *value system estimation* [3, 53], which learns the value systems of its members; and iii) *value system aggregation* [22, 25], which derives a value system to represent the whole society.

Value inference, however, does not cover *value learning* [56]: the problem of eliciting the value alignment of decisions based on human interactions/examples. Value learning implementations remain scarce, though. Exceptions include Anderson et al. [2], who learn ethical principles via inductive logic programming; Wynn et al. [65], who propose representational alignment for learning values in LLMs; and some works using deep learning [38, 61, 62].

Using the previous terminology, we can classify our work as a method for value learning combined with value system *estimation of/aggregation into* distinct groups in a society. The latter distinguishes our approach from the value inference schema, as value system aggregation has so far considered the aggregation of previously elicited value systems into a single one for the society [22, 25], while we estimate separate value systems for different groups.

2.2 Multi-objective RL for Value Learning

For value learning [56], a decision-making framework that explicitly represents the alignment of alternatives with multiple values is required. The most widely used is multi-objective decision-making [55, 58]. In particular, modelling value alignment with utilitarian rewards in (multi-objective) *reinforcement learning*, (MO)RL, is considered a promising approach [11, 21] and has even been used to approximate other ethical theories such as deontology [26].

Value learning in MORL may be tackled via multi-objective inverse reinforcement learning [20, 38] (MOIRL), the problem of learning a reward function for each goal/value and estimating the goal-based behaviour of individuals as the policy that maximizes a combination of these rewards, by observing behaviour traces. However, value learning based on these traces is risky, as the learned rewards might only be able to distinguish the *best* value-aligned behaviours from the rest, failing at evaluating suboptimal choices.

An alternative is *Preference-based Reinforcement Learning* (PbRL) [63] –or RLHF [10]–, the problem of learning a RL policy based on trajectory comparisons. PbRL often consists of two phases: first, a reward model that captures preferences is learned; second, a policy that maximizes this model is obtained via RL. Thus, these methods are a way to perform inverse RL [6]. A single pass through these phases (*offline* PbRL) may misgeneralize after deployment, except under restrictive conditions [68]. As such, *online* PbRL (or human-in-the-loop, HiL) has been suggested, where an agent seeks to maximize positive feedback in interaction with a human using a limited number of queries [32]. Online feedback is particularly desirable for value learning, as it fosters self-reflection [23].

However, single-objective PbRL cannot capture diverse preferences (value systems) or simultaneously represent multiple goals (values). The first limitation has been addressed through personalization and clustering, mainly in the context of generative AI [8, 37, 39, 69]. The second has motivated multi-objective PbRL [31] (PbMORL), either with manually specified objectives [27] or with objectives lacking an explicit AI-usable representation [61].

To our knowledge, our approach is the first to address both limitations simultaneously, albeit with trade-offs. Some authors [37, 39] personalize LLMs via user embeddings; by contrast, our method avoids them, but requires more data for new agents. In [8], a single policy is learned that aggregates a preference-based clustering of agents, an alternative to our per-cluster policy solution. In [61], rewards are learned for several goals and weighed at inference stage to adapt to new users; however, the analysis of user groups is not clarified. In [62], the generative model alignment problem is modelled with transparent value constraints, but requires expensive interactions to elicit the feasible/wanted constraints. Finally, in [31] , theoretical results on Pareto efficiency in PbMORL are provided, but the estimation of particular user preferences is not contemplated.

3 REPRESENTING VALUE SYSTEMS IN MDP

In the following, we adapt and extend our previous model for representing values and value systems in a society [16] to problems involving sequential decisions, and discuss its properties.

3.1 Representing Value Alignment

We set out from a set of m values $V = \{v_1, ..., v_m\}$, where each value v_i constitutes a label for a human value. When *grounded* in a decision-making domain, a value label acquires a particular meaning. Here, we focus specifically on decision-making domains modelled via MDPs. In particular, the meaning of each value label is grounded by the notion of the *alignment* of a set of trajectories with this value, defined through a preference relation. Here, a trajectory τ of length $|\tau| = n$ is a sequence $((s_0, a_0), \ldots, (s_{n-1}, a_{n-1})) \in (S \times A)^n$ where S is the set of states in the MDP and A the available actions.

Definition 3.1 (Value Alignment). The alignment of a set of trajectories $\mathcal{T}$ with a value v_i (in general, the alignment preferences with v_i) is represented by a weak order $\preccurlyeq_{v_i}$ over $\mathcal{T}$, where $\tau \preccurlyeq_{v_i} \tau'$ means that the trajectory τ' is at least as aligned with value v_i as τ.

Inspired by other works in the area [30, 45, 51], we use a specific kind of utility function, in our work called *alignment function*, $\mathcal{A}_{v_i}$ to quantify value alignment, i.e. to represent the relation $\preccurlyeq_{v_i}$: i.e., for all $\tau, \tau' \in \mathcal{T}$: $\tau \preccurlyeq_{v_i} \tau' \iff \mathcal{A}_{v_i}(\tau) \leq \mathcal{A}_{v_i}(\tau')$. Then, to specify the semantics of a set of values, we define the notion of *grounding*.

Definition 3.2 (Grounding). A **grounding** of the set of values V is a set of weak orders $\preccurlyeq_V = \{\preccurlyeq_{v_i}\}_{i=1}^{m}$. Given the respective alignment functions, a **grounding function** for V is: $G_V = (\mathcal{A}_{v_1}, \ldots, \mathcal{A}_{v_m})$.

Given that most behaviours in MDP scenarios are modelled through reward functions, we simplify the set of possible grounding functions to those that can be implemented with a multi-objective reward function vector $\boldsymbol{R} : S \times A \to \mathbb{R}^m$. We say that $\boldsymbol{R}$ implements a grounding function G_V when, for all $\tau \in \mathcal{T}$[2]:

$$G_V(\tau) = \sum_{i=0}^{|\tau|} \boldsymbol{R}_V(s_i, a_i) \tag{1}$$

We write $\boldsymbol{R}(s,a) = (R_{v_1}(s,a), \ldots, R_{v_m}(s,a))$ to denote the reward vector for a given state-action pair (s,a); and for each $v_i \in V : R_{v_i} : S \times A \to \mathbb{R}$ represents the *value reward*, that is, the alignment of (s,a) with value v_i. An MDP with such a reward vector constitutes a Multi-Objective MDP (MOMDP) [58].

3.2 Representing Value Systems

An agent's *value system* expresses the importance assigned to each value. Given a grounding, it induces preferences over trajectories.

Definition 3.3 (Value system). Let $\preccurlyeq_V$ be a grounding for a set of values V. The **value system** of an agent j, based on the grounding $\preccurlyeq_V$, is determined by a weak order $\preccurlyeq_V^j$ over $\mathcal{T}$. If $\tau \preccurlyeq_V^j \tau'$, we say that τ is equally or more aligned than τ with j's value system.

Following other work in the field [22, 25, 45, 59], we assume that the value system of an agent can be expressed through a linear combination of the alignment of a trajectory with the values (Definition 3.4). Our framework can then be seen as *welfarist utilitarian* [54], where values are conceived as different sources of good, and each agent's value system weighs their relative importance.

Definition 3.4 (Value System Function). Let j be an agent with value system $\preccurlyeq_V^j$ and grounding function G_V. The function $\mathcal{A}_{W_j,G_V}(\tau) = W_j \cdot (\mathcal{A}_{v_1}(\tau), \ldots, \mathcal{A}_{v_m}(\tau))^T$ is a **value system function** for j if it represents $\preccurlyeq_V^j$ over $\mathcal{T}$, i.e., for all $\tau, \tau' \in \mathcal{T}$:

$$\mathcal{A}_{W_j,G_V}(\tau) \leq \mathcal{A}_{W_j,G_V}(\tau') \iff \tau \preccurlyeq_V^j \tau'$$

where $W_j = (w_j^{v_1}, \ldots, w_j^{v_m})$ are the *value system weights* that represent the relative importance of each value. We consider normalized weights in the unit $(m-1)$-simplex: $W_j \in (0,1)^m$, $\sum_{i=1}^{m} w_j^{v_i} = 1$.

[2] In Eq. 1, if there is a general concern to prioritize short-term value alignment, we can consider a cumulative discount factor $\gamma < 1$, obtaining $G_V(\tau) = \sum_{i=0}^{|\tau|} \gamma^i \boldsymbol{R}_V(s_i, a_i)$.

If G_V is implemented by a reward vector, we express $\mathcal{A}_{W_j,G_V}$ as:

$$\mathcal{A}_{W_j,G_V}(\tau) = \sum_{i=0}^{|\tau|} W_j \cdot \boldsymbol{R}(s_i, a_i)^T. \tag{2}$$

Thus, the *value system reward* for each agent becomes a linear scalarization of $\boldsymbol{R}$ with weights W_j:

$$R_j(s,a) = W_j \cdot \boldsymbol{R}(s,a)^T$$

We also define the value system represented by R_j as $\preccurlyeq_{\boldsymbol{R}}^{W_j}$.

3.3 Representing the Value System of a Society

Values are inherently social notions [35, 49], and different agents hold different value systems [22]. We set out from a society of agents J, each with its own individual value system $\preccurlyeq_V^j$ based on its individual grounding $\preccurlyeq_{V,j} = (\preccurlyeq_{v_1,j}, \ldots, \preccurlyeq_{v_m,j})$.

Agents can potentially have different views on the meaning of values, i.e. their groundings might differ. However, within human societies and many application domains, a *social* grounding exists [16], i.e. there is a near-consensus on the meaning of values. An example in the real world is the medical domain [44]. A consensual grounding is further necessary as a basis for interpretation and comparison of the value systems of agents. Thus, we assume that such a meaningful social grounding is obtainable, i.e. a set of preference relations $\preccurlyeq_V = \{\preccurlyeq_{v_1}, \ldots, \preccurlyeq_{v_m}\}$ exists that is sufficiently coherent with the variety of individual groundings in the society.

By contrast, we acknowledge that the importance that each agent gives to each value (their value systems) can vary substantially. This is evidenced, for example, in the world values survey [15], where people from different countries hold diverging value-based opinions. Still, since people in the same society have their value system influenced by culture, we can expect regularities in their value systems across social groups [13]. This suggests that in our society, finding subgroups of agents that can be represented by a shared (possibly aggregated [22]) value system is a natural expectation.

The previous considerations lead us to define the value system of a society [16] as the composition of a social grounding with a set of value systems that represent the value-based preferences of different groups of agents in the society.

Definition 3.5 (Value system of a society). A **value system of the society** J (or **social value system**) is a tuple $(\beta, \Omega, \preccurlyeq_V)$ where $\Omega = \{\preccurlyeq_V^l\}_{l=1}^{L}$ is a set of L value systems based on a grounding $\preccurlyeq_V$, and $\beta : J \to \{1, \ldots, L\}$ is a function that assigns each agent to one value system.

We refer to the group of agents assigned, through β, to the l-th value system ($\preccurlyeq_V^l$) as the l-th *cluster* of the society.

4 SOCIAL VALUE SYSTEM LEARNING IN MDPS

In this section, we define the problem of learning the value system of a society of agents based on examples of value alignment and value-based preferences, and present an algorithm for solving it.

4.1 Social Value System Learning Problem

We assume that for each agent j, we have access to examples of their value system and value alignment preferences over pairs of trajectories. In particular, for each agent j there is a dataset

DS_j composed by entries of the type $(\tau, \tau', y_V^j, y_{v_1}^j, \ldots, y_{v_m}^j)$, where $y^j \in \{0, 0.5, 1\}$ indicates as to whether agent j prefers τ over τ' (1), τ' over τ (0) or is indifferent (0.5), with regard to j's value system (y_V^j) and to each of the individual values $(y_{v_i}^j)$. In the sequel, we use D_j to refer to the set of trajectory pairs $\{(\tau, \tau')\}$ that are included in a dataset DS and define $DS = \bigcup_{j \in J} DS_j$ and $D = \bigcup_{j \in J} D_j$.

First , we need to quantify value system and alignment preference differences. To do this, we define the *discordance* between two preference relations by normalizing the number of ordered pairs of trajectories in a set $S \subseteq \mathcal{T} \times \mathcal{T}$ that are ranked differently. Given two relations $\preccurlyeq^1$ and $\preccurlyeq^2$ *discordance* is calculated as follows:

$$d_S\left(\preccurlyeq^1, \preccurlyeq^2\right) = \frac{1}{|S|} \sum_{(\tau,\tau') \in S} \mathbb{1}\left(\left(\tau \preccurlyeq^1 \tau'\right) \not\equiv \left(\tau \preccurlyeq^2 \tau'\right)\right) \tag{3}$$

where $\mathbb{1}\left(\left(\tau \preccurlyeq^1 \tau'\right) \not\equiv \left(\tau \preccurlyeq^2 \tau'\right)\right)$ yields 1 when the preference given by $\preccurlyeq^1$ and $\preccurlyeq^2$ over the pair (τ, τ') differs, and 0 otherwise.

To quantify the degree by which a candidate social grounding represents the individual groundings of a set of agents (i.e. a given society), we employ the notion of *coherence*.

Definition 4.1 (Coherence). Let $\{\preccurlyeq_{v_i,j} \mid j \in J\}$ be the set of value alignment preferences with value v_i held by each agent in a society J. The **coherence** of an alignment preference $\preccurlyeq_{v_i}$ with regard v_i with $\{\preccurlyeq_{v_i,j} \mid j \in J\}$, over the trajectory pairs D is given by:

$$\textsc{chr}_D(\preccurlyeq_{v_i}) = 1 - \frac{1}{|J|} \sum_{j \in J} d_{D_j}(\preccurlyeq_{v_i}, \preccurlyeq_{v_i,j})$$

We define the coherence of a grounding $\Omega = (\preccurlyeq_{v_1}, \ldots, \preccurlyeq_{v_m})$ by $\textsc{chr}_D(\preccurlyeq_V) = \frac{1}{m} \sum_{i=1}^{m} \textsc{chr}_D(\preccurlyeq_{v_i})$.

As we assume that there is a near-consensus in the society concerning value meaning, we assume that a social grounding $\preccurlyeq_V$ exists that has a coherence with the agents' alignment preferences that tends to 1. In a MDP, this suggests that a reward vector $\boldsymbol{R}$ can be learned to represent $\preccurlyeq_V$ with high coherence. We denote such learned grounding with $\preccurlyeq_{\boldsymbol{R}} = (\preccurlyeq_{R_{v_1}}, \ldots, \preccurlyeq_{R_{v_m}})$. Additionally, we can estimate the social value system using the grounding $\preccurlyeq_{\boldsymbol{R}}$ and a set of L value system weights $\boldsymbol{W} = (W_1, \ldots, W_L)$: following Section 3.2, we estimate the value system of the l-th cluster of the society with the value system represented by the reward $R_l(s,a) = W_l \cdot \boldsymbol{R}(s,a)^T$, i.e. for each l: $\preccurlyeq_V^l \approx \preccurlyeq_{\boldsymbol{R}}^{W_l}$. Then, each agent's value system $\preccurlyeq_V^j$ is represented with the reward function of its cluster, i.e. $R_j(s,a) \triangleq R_{\beta(j)}(s,a) = W_{\beta(j)} \cdot \boldsymbol{R}(s,a)^T$. We denote the social value system implemented by $\boldsymbol{R}$ and $\boldsymbol{W}$ with $(\beta, \boldsymbol{W}, \boldsymbol{R})$, to refer to the social value system $(\beta, \Omega, \preccurlyeq_V)$ (Definition 3.5) where $\Omega = \{\preccurlyeq_{\boldsymbol{R}}^{W_l}\}_{l=1}^{L}$ and $\preccurlyeq_V = \preccurlyeq_{\boldsymbol{R}}$.

To evaluate the quality of a value system of the society we use two metrics from our previous work [16]: *representativeness* (Definition 4.2) and *conciseness* (Definition 4.3). The first refers to the accuracy with which each agent's value-based preferences are recovered by their assigned value system. The latter measures the differences between the value systems represented in the solution, which motivates the use of fewer clusters to describe the society.

Definition 4.2 (Representativeness of a value system of a society). The **representativeness** of a social value system $(\beta, \Omega, \preccurlyeq_V)$ over the trajectory pairs $D = \bigcup_{j \in J} D_j$ is:

$$\textsc{repr}_D(\beta, \Omega, \preccurlyeq_V) = 1 - \frac{1}{|J|} \sum_{j \in J} d_{D_j}\left(\preccurlyeq_V^{\beta(j)}, \preccurlyeq_V^j,\right)$$

Definition 4.3 (Conciseness of value system of a society). The **conciseness** of a social value system $(\beta, \Omega, \preccurlyeq_V)$ over D is:

$$\textsc{conc}_D(\beta, \Omega, \preccurlyeq_V) = \min_{l \neq l'} d_D(\preccurlyeq_V^l, \preccurlyeq_V^{l'}),$$

Both metrics are normalized in $[0, 1]$ and higher values indicate a better representation of the agent's value systems, and higher significance of the clusters composing the social value system (in terms of their ability to represent different preferences), respectively.

To tackle the trade-off between conciseness and representativeness, we propose minimizing a heuristic metric $\Gamma(\textsc{repr}, \textsc{conc})$. Natural choices for Γ are clustering metrics that compare inter-cluster and intra-cluster distances, such as the Dunn-Index [12] or the Ray-Turi Index [42]. Here, we propose using a version of the latter, as it can simply be expressed by the ratio $(1 - \textsc{repr})/\textsc{conc}$, i.e. the mean of the distances from each agent to the centroids (the value systems in each cluster) over the minimum distance among centroids. As conciseness is likely to reach 0, and to prioritize representativeness over conciseness, we use $\Gamma(\textsc{repr}, \textsc{conc}) = (1 - \textsc{repr})/(1 + \textsc{conc})$. The analysis of alternative Γ metrics is left for future work.

We are now in a position to define the *social value system learning problem*, as learning a value system of the society $(\beta^*, \boldsymbol{W}^*, \boldsymbol{R}^*)$ that solves the following bi-level optimization problem:

$$\begin{aligned} (\beta^*, \boldsymbol{W}^*) \in \arg\min_{\beta, \boldsymbol{W}} \Gamma\left(\textsc{repr}_D(\beta, \boldsymbol{W}, \boldsymbol{R}^*), \textsc{conc}_D(\beta, \boldsymbol{W}, \boldsymbol{R}^*)\right) \\ \text{subject to} \quad \boldsymbol{R}^* \in \arg\max_{\boldsymbol{R}} \textsc{chr}_D(\preccurlyeq_{\boldsymbol{R}}) \end{aligned} \tag{4}$$

Problem 4 formulates the aim of maximizing the trade-off between representativeness and conciseness, based on finding a grounding with maximal coherence. This ensures that the learned value system weights are based on a correct estimation of the social grounding.

4.2 Learning Value-aligned Behaviours with HiL

The learned value systems for each cluster should represent well the corresponding agents' behaviours. In MDPs, different behaviours are represented via *policies* that obtain an accumulated reward. Although human agents may occasionally be misaligned, value-aware software agents should remain consistently value-aligned. Thus, they should follow *rational* value-aligned policies, i.e. those that maximize an accumulated reward that takes into account the grounding of the values. Assuming that all values are worth being promoted, the (software) agents' policies are, further, expected to be Pareto-efficient regarding this grounding.

Given a cluster l described by a reward vector $\boldsymbol{R}$ and weights W_l, we can define a policy π_l that maximizes $R_l(s,a) = W_l \cdot \boldsymbol{R}(s,a)$. This policy should be *aligned* with the behaviours of the agents belonging to the cluster l, and Pareto-efficient regarding the social grounding that $\boldsymbol{R}$ approximates.

However, in our setting, we learn the society's value system over a finite dataset of trajectory comparisons of unknown nature. Even if we find cluster value systems with high representativeness, they may not represent well the actual agents' behaviours. The reason is

that the learned rewards do not always *generalize* well to the MDP dynamics. Capturing the agents' choices from generic comparisons might not be enough. Instead, we should try to acquire value systems that not only correctly evaluate suboptimal trajectories, but also tend to prefer the ones that are close to the agents' behaviours.

To address the last issue, inspired by human-in-the-loop (HiL) PbRL [32], we propose asking for value-based and alignment preferences over trajectories sampled from estimates of the policies π_l while learning a social value system. This "online" feedback should steer the learned policies towards the actual agent behaviours.

4.3 Algorithm for Social Value System Learning

We employ a deep learning solution approach to Problem 4. We use two models. First, a reward vector network $\boldsymbol{R}^\theta$ with parameters θ, that represents a social grounding. Second, L_{max} neural networks. Each network consists of a linear layer given by certain value system weights W_l^ω that are parametrized with $\omega \in [\mathbb{R}^m]^{L_{max}}$ through a *softmax* calculation: $W_l^\omega = (w_j^{v_1}, \ldots, w_j^{v_m}) = \frac{\exp \omega_l}{\sum \exp(\omega_l)}$, so they remain in the simplex. We then write $\boldsymbol{W}^\omega = (W_1^\omega, \ldots, W_{L_{max}}^\omega)$, but given an assignment β, we only consider the necessary $L \leq L_{max}$ networks/weights. Then, the value system estimated through these networks would be written as $(\beta, \boldsymbol{W}^\omega, \boldsymbol{R}^\theta)$.

For learning purposes, we represent the grounding preferences and the value-based preferences induced by these networks in the datasets using the differentiable Bradley-Terry (BT) model [5]. It estimates the relative preference between two trajectories given a reward function (Eq. 5)[3]. With this model, we consider $p(\tau, \tau'|R) = 0.5$ only if τ and τ' are similarly aligned and it tends to 1 or 0 if their difference in alignment is increasingly strict. We then use cross-entropy-like loss calculations to reproduce the discordance of our BT model with the dataset preferences. Finally, we use gradient descent in a structured way to update the networks $\boldsymbol{W}^\omega, \boldsymbol{R}^\theta$ in line with the goals in Problem 4 (details in supplementary material [17]).

$$p(\tau, \tau'|R) = \frac{\exp \sum_{(s,a)\in\tau} R(s,a)}{\exp \sum_{(s,a)\in\tau} R(s,a) + \exp \sum_{(s',a')\in\tau'} R(s',a')} \quad (5)$$

Algorithm 1 is our proposed, approximate solution to Problem 4. It builds upon our previous work [16] (which solves Problem 4 in non-sequential decision making) and PbMORL [31], a MORL algorithm that estimates a set of Pareto-efficient policies from online preferences. It uses Envelope Q-Learning [67] (EQL) to learn a multi-objective policy conditioned on weights: $\Pi(s, a|W)$. This allows to request for the policy π_l that maximizes $W_l \cdot \boldsymbol{R}$ with $\pi_l = \Pi(s, a|W_l)$. By sampling different weights we get approximately efficient policies that represent the agents' value systems. Our system introduces key changes, as its goal is not finding the whole front, but rather, a smaller set of weights and policies that represent the agents' values.

Initialization. Our algorithm (SVSL-P, Algorithm 1) starts by finding an approximation to Problem 4 using a static dataset DS, to have an initial "good guess" of the reward vector and the clusters prior to learning the RL policies. To do so, we execute a version of Algorithm 2 from [16] (Algorithm S2 in the supplementary material [17]). The algorithm performs *Expectation Maximization* (see below) with occasional random mutations to avoid local convergence problems [64], and it is run until we reach a certain predefined score $A_{ref} \in [0, 1]$ in both grounding coherence and representativeness.

Main loop (Lines 3-17) Now, the aim is to iteratively improve the solution (the social value system $(\beta, \boldsymbol{W}^\omega, \boldsymbol{R}^\theta)$) (*Value system update*) while learning the corresponding MO policy $\Pi(s, a|W)$ (*Policy update*) using HiL with environment experiences (*Exploration*).

Exploration (Lines 3-5). At every timestep (main loop iteration) we collect a transition $(s, a, \boldsymbol{R}^\theta(s, a), s', W)$ and add it to an *experience* replay buffer R_e. The action is selected by an ϵ-greedy version of the current policy estimate $\Pi(s, a|W)$ supplying a randomly selected weight combination W from those in $\boldsymbol{W}^\omega$.

Value system update (Lines 6-12). Every K exploration timesteps, we perform two consecutive tasks:

(1) *Preference collection* (Lines 7-8). We extract new preference feedback by asking a random group of N_a agents about a set of N_s pairs of trajectories in R_e selected with *QPA* [19]. Namely, we ask each agent for its value alignment and value-based preferences about each trajectory pair, and this data is inserted in a *preference* buffer R_p. This step is present in PbMORL, but that algorithm can ask for preferences derived from any weight combination in the whole simplex. The feedback required by our solution is far more simple: an agent is only asked to answer about value alignment and its own value system.

(2) *Expectation-Maximization (EM)* (Lines 10-11). Similar to PbMORL, here we improve the reward vector estimate $\boldsymbol{R}_V^\theta$. In our case, we also update the value system weights $\boldsymbol{W}^\omega$ and revise the clustering (updating L and β), by performing a series of E_r EM cycles. We take b_{ep} random entries from R_p for each agent; add them to the static dataset DS; and perform a "hard" **E-step**: the assignment of each agent into the cluster that best represents its value system. Then, we perform the **M-step** (m_r times) to update $\boldsymbol{R}^\theta$ and $\boldsymbol{W}^\omega$ based on a random batch of entries of size b_{mp} in R_p. After E_r EM cycles[4], we obtain a new social value system estimation $(\beta, \boldsymbol{W}^\omega, \boldsymbol{R}^\theta)$. We then use $\boldsymbol{R}^\theta$ to update the experience buffer rewards in R_e.

Policy update (Lines 13-16). We apply the EQL Q-learning step from PbMORL so that $\Pi(s, a|W)$ learns to maximize the scalarization of $\boldsymbol{R}^\theta$ under weights and transitions collected from b_π experiences in R_e, selected via prioritized *hybrid experience replay* [19][5].

After a number of timesteps T, the algorithm returns a social value system $(\beta, \boldsymbol{W}^\omega, \boldsymbol{R}^\theta)$ and the multi-objective policy $\Pi(s, a|W)$ that models the policies that represent each cluster: $\pi_l = \Pi(s, a|W_l)$.

5 EVALUATION

We evaluate our proposal in two synthetic MDPs. In each of them, we simulate a society of agents J that considers a social grounding implemented by a "ground truth" reward vector $\boldsymbol{R}$, and have a value system determined by specific weights $\{W_j | j \in J\}$. We selected the weights and generated the static dataset DS as follows. We trained EQL [67] to convergence in each case (see parameters in

[3] In Eq. 5, the designer might consider a discount factor $\gamma < 1$ as in Eq. 1.

[4] See supplementary material [17] for a complete description of the EM process.

[5] In particular, we collect $b_\pi/2$ transitions randomly among the most recent ones, and the rest are selected through prioritized experience replay [48].

[6] The parameter λ referes to the Lagrange multipliers used when updating $\boldsymbol{W}^\omega$ and $\boldsymbol{R}^\theta$ to prioritize coherence over representativeness (See supplementary material [17]).

Algorithm 1 Social Value System Learning in MDP (**SVSL-P**)

Input: Dataset DS. Maximum number of clusters L_{max}. Reference accuracy A_{ref}. Timesteps between VS updates K. Total timesteps T. Trajectory sampling size N_s. Asked agents per VS update N_a. Per-agent batch size b_{ep} (for E-steps) and b_{mp} (for M-steps). EQL batch size b_π and gradient steps T_π. Parameters of Algorithm S1 and S2 in supp. material.

Output: Assignment of agents into clusters β, a reward vector $\boldsymbol{R}^\theta$, value systems $\{W_l^\omega\}_{l=1}^L$ and weight-conditioned policy Π.

Initialize replay buffers R_e, R_p and a Q network $Q(s,a|W)$.
$\beta_0, \boldsymbol{R}^\theta, \boldsymbol{W}^\omega, \lambda \leftarrow$ Run **Algorithm S2**, supp. material, until $\min_i \text{CHR}_D \geq A_{ref}$, $\text{REPR}_D \geq A_{ref}$. [6]
for environment timestep $t = 0, \ldots, T-1$ **do**
  Sample one set of VS weights w from those in $\boldsymbol{W}^\omega$.
  Collect transition $(s, a, R_V^\theta(s,a), s', d, w)$, add it to R_e
  if $t \mod K = 0$ **then**
    $B_a \leftarrow$ Select N_a agents at random (no replacement).
    $B_p \leftarrow$ Collect N_s pairs of trajectories from R_e (**QPA**).
    Preference collection: add new $B_a \times B_p$ entries into R_p.
    $\beta, \theta, \omega, \lambda' \leftarrow$ *Expectation-Maximization (EM)*: **Algorithm S1**($\beta, \theta, \omega, \lambda, R_p, b_{ep}, b_{mp}$), supp. material.
    Update $\boldsymbol{R}^\theta$ and relabel R_e with it. Update $\boldsymbol{W}^\omega$.
  end if
  for gradient step in T_π **do**
    Sample a minibatch from replay buffer R_e
    Apply the *update* step from the EQL paper [67] on $Q(s,a|W)$ using a batch of size b_π from R_e as in [31].
  end for
end for
return ($\beta, \boldsymbol{W}^\omega, \boldsymbol{R}^\theta$), $\Pi(s,a|W)$ (from Q).

the supplementary material [17]). Then, we sampled 50 equally spaced weight combinations in the weight space. We executed the learned policy for each weight ($\Pi(s,a|w)$) in the environment to produce 1000 trajectories. We selected a set of M_W value system weight combinations whose policies form all the points on the Pareto front found by EQL. We created M_A agents per value system, for a total of $M_A \cdot M_W$ agents, and sampled 200 trajectories per agent. A proportion of these trajectories (r_p) are "rational": they are obtained from the ϵ-greedy application of the policy obtained by EQL using the agent's weights. The rest are sampled via a random policy. The selection of ϵ and r_p varies in environments to get trajectories of varied quality. From these trajectories, per agent, we sampled 200 value alignment (per value) and 200 value-based preference pairs at random. Then, 50% of the comparisons of each agent are used to form the static dataset DS, the rest are left out as a test dataset.

In each environment, we compare the results obtained by two baselines, a state-of-the-art algorithm and our method (**SVSL-P**).

The first baseline is **EQL** [67] with the original reward vector $\boldsymbol{R}$. We use the execution that generated the agents and datasets. The second baseline is the social value system learning algorithm from [16] (Algorithm S2 in the supplementary material [17], **SVSL**), a preference-based reward learning and clustering baseline (without HiL). To learn the associated policies for the obtained value systems, we run EQL with the learned reward vector with SVSL.

The state-of-the-art related algorithm that we analyse is **Pb-MORL** [31]. It learns a reward vector and a weight-dependent policy $\Pi(s,a|W)$. Since the algorithm does not extract clusters, we first select $|J|$ equally spaced weight combinations in the simplex to represent potential cluster value systems. Then, we use the E-step from Algorithm S1 to assign each agent to the cluster that best represents its value system and discard the empty ones.

We run the algorithms 10 times (different seeds) with the same datasets. The maximum number of clusters for SVSL and SVSL-P was set to $L_{max} = 10$ and $L_{max} = 15$ for the FF and the MVC environments, respectively. The hyperparameter specifications for all algorithms are detailed in the supplementary material [17].

5.1 Firefighters Environment

In the firefighters environment (FF), adapted from [36] and tested in previous works [18], different agents (firefighters) are trained to rescue people and put out a fire in a building, remaining aligned with two values: *professionalism* (p_f) and *proximity* (p_x). The first promotes behaving under firefighters' best practices, while the latter promotes actions with the goal of saving lives at all costs. The environment consists of 5 actions and 400 possible states with ground-truth value rewards bounded in $[-1, 1]$. The environment specification details are available in the supplementary material [17].

The reward vector model observes a one hot-encoded version of the state-action features (in supplementary material). We use a neural network for each value. Each one consists of three fully-connected hidden linear layers (128 neurons each) followed by an output layer (one neuron, no bias) and $Tanh$ activation functions. We use the same network configuration for all algorithms.

Using EQL we obtained $M_W = 5$ value systems whose policies form the convex part of the Pareto front. We simulated $M_A = 3$ agents per value system and sampled 200 trajectories with each policy. The proportion of rational trajectories was set to $r_p = 80\%$, and those were sampled using ϵ-greedy policies with $\epsilon = 0.1$.

5.2 Multivalued Car Environment

The Multivalued Car Environment (MVC) was proposed in [45] to illustrate the problem of learning norm-abiding and value-aligned policies given a computational value specification. In MVC, a car agent wants to reach a destination (value of *achievement*) in a road grid while promoting the value of *safety* by respecting pedestrians and avoiding bumpy areas for the sake of *comfort*.

We use the same neural network architecture per value as in FF to estimate the reward vectors. The final layer of each network is left without an activation function (thus, also, unbounded). To avoid overflows, we clamp the learned rewards in the interval $[-100, 100]$. Also, because the environment episodes are of very different lengths, we introduced a discount factor $\gamma = 0.99$ to calculate the alignment of the trajectories with values/value systems.

We obtained $M_W = 14$ value systems and Pareto-optimal policies with EQL. We created $M_A = 2$ agents per value system in the front. To build DS, we sampled 200 trajectories with $r_p = 80\%$, and $\epsilon = 0.1$.

Method (in FF)	L (histogram)	REPR ↑	Prof. CHR ↑	Prox. CHR ↑	CONC ↑	Ray-Turi ↓	PF size (all) ∼	PF size (cls.) ∼	HV. (all) ↑	HV (cls.) ↑	MUL ↓ (all)	MUL ↓ (cls.)
EQL	-	-	-	-	-	-	5	5	40.52	40.52	0.0	0.0
SVSL		0.915 ±0.022	0.860 ±0.042	0.858 ±0.019	**0.163** ±0.055	0.074 ±0.024	1.1 ±0.3	2.0 ±0.0	0.0 ±0.0	0.0 ±0.0	45.118 ±0.95	46.653 ±1.314
PbMORL		0.922 ±0.008	0.863 ±0.009	0.920 ±0.012	0.067 ±0.047	0.074 ±0.010	3.8 ±0.600	2.4 ±0.489	35.35 ±0.934	31.39 ±2.98	0.670 ±0.154	1.240 ±0.579
SVSL-P		**0.968** ±0.010	**0.967** ±0.014	**0.951** ±0.009	0.049 ±0.016	**0.03** ±0.010	3.6 ±0.489	3.8 ±0.400	**38.98** ±0.170	**38.83** ±1.454	**0.117** ±0.150	**0.174** ±0.293

Method (in MVC)	L (histogram)	REPR ↑	Ach. CHR ↑	Safe. CHR ↑	Comf. CHR ↑	CONC ↑	Ray-Turi ↓	PF size (all) ∼	PF size (cls.) ∼	HV (all) ↑	HV (cls.) ↑	MUL ↓ (all)	MUL ↓ (cls.)
EQL	-	-	-	-	-	-	-	14	14	1.233	1.233	0.0	0.0
SVSL		0.771 ±0.112	0.775 ±0.134	0.682 ±0.139	0.628 ±0.165	**0.051** ±0.031	0.139 ±0.053	4.4 ±1.43	2.0 ±0.0	0.556 ±0.510	0.475 ±0.462	14.763 ±11.24	18.562 ±12.06
PbMORL		0.890 ±0.013	0.884 ±0.038	0.750 ±0.044	0.649 ±0.043	0.031 ±0.006	0.130 ±0.014	15.6 ±2.8	7.2 ±1.167	**1.217** ±0.011	**1.191** ±0.024	**1.102** ±0.387	1.784 ±0.668
SVSL-P		**0.908** ±0.013	**0.886** ±0.015	**0.883** ±0.009	**0.748** ±0.023	0.044 ±0.012	**0.088** ±0.013	8.9 ±1.375	5.0 ±0.775	1.197 ±0.017	1.175 ±0.026	1.152 ±0.542	**1.522** ±0.644

Table 1: FF (top) MVC (bottom). For 10 random seeds, number of clusters obtained and frequency (L), representativeness (REPR), coherence for each value (CHR), conciseness (CONC), and Ray-Turi index running each algorithm with 10 random seeds over the test-sets. In the right part of the charts, cardinality (PF size), hypervolume (HV) and Maximum Utility Loss (MUL) of the obtained Pareto fronts using all candidate value system weights ("all" variant); and of the Pareto front obtained using only the learned weights for the value system of the societies ("cls." variant). Results indicate the average values and standard deviations.

5.3 Discussion of results

We first discuss the value systems of the society learned by SVSL, PbMORL and SVSL-P in terms of number of clusters, representativeness, coherence and conciseness over the test set. We focus on the left part of the charts in Table 1 (FF at the top, MVC at the bottom).

SVSL obtains relatively low representativeness (accuracy in representing the individual value systems of agents) and grounding coherence (accuracy of the learned grounding to represent the simulated one) in the test sets, which implies that it did not generalize to the test dataset. PbMORL and SVSL-P generalized better in terms of representativeness, at the expense of conciseness. SVSL-P in particular, outperforms PbMORL in terms of grounding coherence, as it focuses explicitly on this aspect of the optimization. SVSL achieves, though, the highest conciseness (relevancy of the cluster value systems learned), but tends, wrongly, to obtain less clusters than required. Notably, in the FF case, SVSL-P achieves a higher representativeness with a similar conciseness than PbMORL while using more clusters, indicating that SVSL-P better recognises the value system diversity. In the MVC domain, though, PbMORL selects significantly more clusters than SVSL-P , but this does not result in better representativeness for PbMORL. The Ray-Turi index quantifies the trade-off between representativeness and conciseness: it is smaller in SVSL-P, confirming its best performance.

We now test whether the learned reward vectors induce policies that are aligned with the value systems of the simulated agents. Since the simulated agents follow the policies that form the convex part of the Pareto front with respect the "ground-truth" grounding, we assess how well the learned policies approximate this front (even if using clusters). The right part of the charts in Table 1, shows quantitative results on the approximated Pareto fronts across methods and environments. We analyse two fronts. The first one is composed by the policies in $\Pi(s, a|W)$ obtained with the weights that each algorithm considers as potential clusters. For SVSL-P and SVSL, these are the value system weights $\boldsymbol{W}^{\omega}$, and for PbMORL, the set of equally spaced weight combinations of size $|J|$ (15 in FF, 28 in MVC). The second front is composed only by the policies learned for the weights finally selected as society clusters. The metrics for the first front are labelled as ("all"), and for the second as ("cls.").

We use three metrics in Table 1 to describe the fronts: *cardinality* (PF size), *hypervolume* (HV) and *maximum utility loss* (MUL) [43]. HV is computed from reference points $(0, 0)$ in FF and $(-40.0, -50.0, -50.0)$ in MVC (HV is scaled down by 10^5 in MVC). MUL is measured against the ground-truth front in FF, and against the front learned with EQL in MVC, as the true front is unknown. Lower MUL indicates closer convergence to these fronts. Higher HV values capture both better convergence and distribution of the cluster policies across the learned front.

SVSL produces incompetent policies in both MDPs, as its learned reward vectors failed to generalize to their dynamics. In contrast, PbMORL and SVSL-P achieve HV and MUL comparable to EQL when considering the full fronts, impressively, despite the bigger size of the PbMORL fronts. Notably, considering the front obtained by the learned clusters alone, the performance in terms of HV of both algorithms is similar, and SVSL-P achieved a better MUL. This occurs even in MVC, where the front of SVSL-P has a smaller size. In FF, though, its size is bigger: but this does not indicate a bad

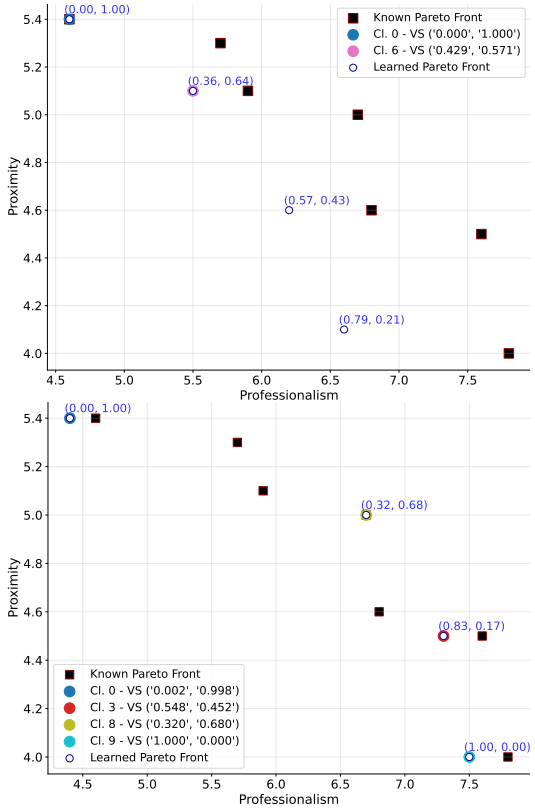

Figure 1: FF environment. Approximated Pareto front and clusters learned with PbMORL (Top) and SVSL-P (bottom, ours) with a particular seed. Black squares form the ground-truth Pareto front. White dots depict weights which policies are in the approximated front. Coloured dots indicate the policies representing each learned cluster (in the legend).

clustering performance, since HV and MUL metrics outperform PbMORL's, showing SVSL-P clusters are of higher significance.

Figure 1 represents the Pareto fronts of PbMORL and SVSL-P estimated from their respective potential value system weights in the FF domain, as well as the evaluation of the policies obtained for each used cluster for a certain seed[7]. Both algorithms approximate well the behaviours in the original front. In both cases, the learned clusters tend to be Pareto optimal for both algorithms, but their distribution across the front is worse for PbMORL. PbMORL typically used only two clusters (Table 1) to represent the agents' value-based preferences, which means the reward vector was not learned in a sufficiently detailed manner to differentiate them properly.

There are also limitations to analyse. First, there are cases where SVSL-P learns two value systems that represent distinct preferences, yet they induce the same policy (supplementary material [17], Figures 2,3). Second, the policies learned were generally not exactly in the ground-truth front (obtained by EQL), despite being "close" (shown by the MUL metric). In the FF domain, for example, the proportion of cluster weights whose policies were in the ground-truth front was, on average, 47.1% for SVSL-P and 40.8% for PbMORL. Lastly, the standard deviations in the front-related metrics (Table 1, right) are high for PbMORL and SVSL-P. This implies that there is some instability across seeds. However, the number of learned clusters is more stable with SVSL-P. Further analysis on these limitations is available in the supplementary material.

[7]We omitted the fronts in MVC given their 3D visualization complexity. The corresponding graph for every seed is available in the supplementary material [17].

6 CONCLUSIONS

We proposed a computational model of the value systems of a society of agents in the context of Markov Decision Processes (MDP). Given a set of human values, we grounded value alignment in a particular MDP with a multi-objective reward vector and represented the value systems of different subgroups of agents (clusters) via linearly scalarized reward functions. We put forward an algorithm that learns an instance of this model from online pairwise trajectory comparisons that are provided by each agent based on both i) its understanding of value alignment (with each value) and ii) the agent's value system. In parallel, it learns an approximately Pareto-efficient MDP policy for each cluster (in terms of value alignment) that represents behaviours aligned with the value systems of its members. Our algorithm, SVSL-P, is based on a Preference-based MORL (PbMORL) method [31] and a previous clustering approach [16].

The results in two synthetic environments show that SVSL-P can learn a concise, representative, and coherent set of value systems to describe a society, and the associated value-aligned behaviours. The learned models generalize in the environment, leading to approximately Pareto efficient policies in terms of the simulated value alignment specifications. The Pareto fronts are focused on the value systems of the agents, and are equally competent to the fronts derived by PbMORL. Our algorithm also requires less intensive human feedback: SVSL-P only asks each agent about their own value systems and their understanding of values. This last advantage is crucial to the applicability of this type of research in real use cases.

There are limitations and avenues for future work. First, the variability across runs should be reduced. Second, the learned policies are not always Pareto efficient regarding value alignment for all preference-based clusters. Third, reducing the number of online queries is needed for real-world applications. For future work, we will revise our methodology to control the obtention of more concise versus representative solutions. Although we have tested this framework with real-world data in non-sequential decision making [16], to broaden the applicability to more realistic scenarios, we finally suggest researching on modelling non-linear and context-dependent value systems (e.g. by selecting distinct value systems under identifiable conditions) and agent-based groundings/rewards.

ACKNOWLEDGMENTS

This work has been supported by grant COSASS: PID2021-123673OB-C32 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", and by project grant EVASAI: PID2024-158227NB-C32 funded by MICIU/AEI/10.13039/501100011033/FEDER, UE. Andrés Holgado-Sánchez has received funding by grant "Contratos Predoctorales de Personal Investigador en Formación en Departamentos de la Universidad Rey Juan Carlos (C1 PREDOC 2025)", funded by Universidad Rey Juan Carlos.

REFERENCES

[1] Alba Aguilera, Nieves Montes, Georgina Curto, Carles Sierra, and Nardine Osman. 2024. Can Poverty Be Reduced by Acting on Discrimination? An Agent-based Model for Policy Making. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 22–30.

[2] Michael Anderson and Susan Leigh Anderson. 2018. GenEth: A general ethical dilemma analyzer. *Paladyn* 9 (2 2018), 337–357. Issue 1. https://doi.org/10.1515/PJBR-2018-0024

[3] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems* 191 (2020), 105184. https://doi.org/10.1016/j.knosys.2019.105184

[4] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2019. Incorporating Behavioral Constraints in Online AI Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 3–11. https://doi.org/10.1609/aaai.v33i01.33013

[5] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. https://doi.org/10.2307/2334029

[6] Daniel S. Brown, Wonjoon Goo, and Scott Niekum. 2019. Better-than-Demonstrator Imitation Learning via Automatically-Ranked Demonstrations. https://arxiv.org/abs/1907.03976

[7] José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics* 26 (2020), 501–532. Issue 2. https://doi.org/10.1007/s11948-019-00151-x

[8] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Proc. 41st Int. Conf. on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. JMLR.org, https://www.jmlr.org, 6116–6135.

[9] Remy Chaput, Laetitia Matignon, and Mathieu Guillermin. 2023. Learning to identify and settle dilemmas through contextual user preferences. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*. IEEE, 474 – 479. https://doi.org/10.1109/ICTAI59109.2023.00075

[10] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[11] Virginia Dignum. 2017. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 1101 Pennsylvania Ave, NW, Suite 300, Washington, DC 20004, 4698–4704. https://doi.org/10.1007/978-3-030-30371-6

[12] J. C. Dunn. 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* 4, 1 (1974), 95–104. https://doi.org/10.1080/01969727408546059 arXiv:https://doi.org/10.1080/01969727408546059

[13] Michael Grenfell. 2014. *Pierre Bourdieu: key concepts*. Routledge.

[14] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3916–3924.

[15] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Milena Lagos, Pippa Norris, Eduard Ponarin, and Bianca Puranen. 2022. *World Values Survey: Round Seven – Country-Pooled Datafile*. Madrid. Version 5.0.0.

[16] Andrés Holgado-Sánchez, Holger Billhardt, Sascha Ossowski, and Sara Degli-Esposti. 2025. Learning the Value Systems of Societies from Preferences. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025) (Frontiers in Artificial Intelligence and Applications, Vol. 413)*. IOS Press, 1123–1130. https://doi.org/10.3233/FAIA250923

[17] Andrés Holgado-Sánchez, Peter Vamplew, Richard Dazeley, Sascha Ossowski, and Holger Billhardt. 2026. Learning the Value Systems of Societies with Preference-based Multi-objective Reinforcement Learning. https://arxiv.org/abs/2602.08835 Full version and supplementary material.

[18] Andrés Holgado-Sánchez, Holger Billhardt, Alberto Fernández, and Sascha Ossowski. 2026. Learning the value systems of agents with preference-based and inverse reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 40, 4 (2026), 4. Issue 1. https://doi.org/10.1007/s10458-026-09732-0

[19] Xiao Hu, Jianxiong Li, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. 2024. Query-Policy Misalignment in Preference-Based Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*. Curran Associates Inc., Red Hook, NY, USA, 28899–28923. https://openreview.net/forum?id=UoBymIwPJR

[20] Daiko Kishikawa and Sachiyo Arai. 2022. Multi-Objective Deep Inverse Reinforcement Learning through Direct Weights and Rewards Estimation. In *2022 61st Annual Conference of the Society of Instrument and Control Engineers (SICE)*. Institute of Electrical and Electronics Engineers (IEEE), 122–127. https://doi.org/10.23919/SICE56594.2022.9905799

[21] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. arXiv:1811.07871 [cs.LG] https://arxiv.org/abs/1811.07871

[22] Roger Xavier Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite López-Sánchez, Pradeep K. Murukannaiah, Juan A. Rodríguez-Aguilar, and Francisco Salas-Molina. 2024. Aggregating value systems for decision support. *Knowledge-Based Systems* 287 (2024), 111453. https://doi.org/10.1016/j.knosys.2024.111453

[23] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel IJ Dobbe, Catholijn M Jonker, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Pradeep K Murukannaiah. 2023. Value Inference in Sociotechnical Systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1774–1780.

[24] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. 2022. What values should an agent align with?: An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems* 36 (2022), 1–32. Issue 1. https://doi.org/10.1007/s10458-022-09550-0

[25] Aarón López-García. 2024. A Proposal for Selecting the Most Value-Aligned Preferences in Decision-Making Using Agreement Solutions. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 1: EAA*. INSTICC, SciTePress, Avenida de S. Francisco Xavier, Lote 7 Cv. C, 2900-616 Setúbal, Portugal, 461 – 470. https://doi.org/10.5220/0012586300003636

[26] Ignacio D. López-Miguel, Sebastian Adam, Ezio Bartocci, Thomas Eiter, and Martin Tappler. 2025. OFTEN-DeepRL: On-the-Fly Teaching of Ethical Norms to Deep Reinforcement Learning Agents. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025) (Frontiers in Artificial Intelligence and Applications, Vol. 413)*. IOS Press, 3315–3322. https://doi.org/10.3233/FAIA251200

[27] Junlin Lu, Patrick Mannion, and Karl Mason. 2024. Inferring preferences from demonstrations in multi-objective reinforcement learning. *Neural Computing and Applications* 36, 36 (12 2024), 22845–22865.

[28] Dhruv Malik, Malayandi Palaniappan, Jaime Fisac, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. 2018. An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). JMLR.org, https://www.jmlr.org, 3394–3402. https://proceedings.mlr.press/v80/malik18a.html

[29] Nieves Montes, Nardine Osman, Carles Sierra, and Marija Slavkovik. 2023. Value Engineering for Autonomous Agents. https://doi.org/10.48550/arXiv.2302.08759 arXiv:2302.08759

[30] Nieves Montes and Carles Sierra. 2022. Synthesis and properties of optimally value-aligned normative systems. *Journal of Artificial Intelligence Research* 74 (2022), 1739–1774. https://doi.org/10.1613/jair.1.13487

[31] Ni Mu, Yao Luan, and Qing Shan Jia. 2024. Preference-based Multi-Objective Reinforcement Learning with Explicit Reward Modeling. In *Proceedings - 2024 China Automation Congress, CAC 2024*. Institute of Electrical and Electronics Engineers Inc., 4874–4879. https://doi.org/10.1109/CAC63892.2024.10865310

[32] Calarina Muslimani and Matthew E. Taylor. 2024. Leveraging Sub-Optimal Data for Human-in-the-Loop Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2399–2401.

[33] Emery A. Neufeld. 2022. Reinforcement Learning Guided by Provable Normative Compliance. In *International Conference on Agents and Artificial Intelligence*, A. Rocha, L. Steels, and J. van den Herik (Eds.), Vol. 3. Science and Technology Publications, Lda, Avenida de S. Francisco Xavier, Lote 7 Cv. C, 2900-616 Setúbal, Portugal, 444 – 453. https://doi.org/10.5220/0010835600003116 Cited by: 3; All Open Access; Green Accepted Open Access; Green Open Access; Hybrid Gold Open Access.

[34] Emery A. Neufeld. 2024. Learning Normative Behaviour Through Automated Theorem Proving. *KI - Kunstliche Intelligenz* 38, 1-2 (2024), 25 – 43. https://doi.org/10.1007/s13218-024-00844-x Cited by: 3; All Open Access; Hybrid Gold Open Access.

[35] Nardine Osman and Mark d'Inverno. 2024. A Computational Framework of Human Values. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1531–1539.

[36] Nardine Osman, Manel Rodríguez-Soto, and Jordi Sabater-Mir. 2025. Instilling Organisational Values in Firefighters through Simulation-Based Training. arXiv:2512.13737 [cs.CY] https://arxiv.org/abs/2512.13737

[37] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E. Ozdaglar. 2024. RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*. OpenReview, https://openreview.net. https://openreview.

net/forum?id=8XI7AGByAp

[38] Markus Peschl, Arkady Zgonnikov, Frans A. Oliehoek, and Luciano C. Siebert. 2022. MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1038–1046.

[39] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24, Vol. 37)*, A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak, and C Zhang (Eds.). Curran Associates Inc., Red Hook, NY, USA, Article 1664, 29 pages. https://proceedings.neurips.cc/paper_files/paper/2024/file/5e1c255653eb98cef13f45b2d337c882-Paper-Conference.pdf

[40] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn Jonker. 2012. Elicitation of situated values: Need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology* 14 (2012), 285 – 303. Issue 4. https://doi.org/10.1007/s10676-011-9282-6

[41] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2338, 14 pages.

[42] Sid Ray and Rose H Turi. 2000. Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. In *4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*. Narosa Publishing House, India, 137 – 143.

[43] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. 2022. Pareto Conditioned Networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1110–1118.

[44] Manel Rodríguez-Soto, Nardine Osman, Carles Sierra, Nieves Montes, Jordi Martinez Roldan, Rocio Cintas García, Cristina Farriols Danes, Montserrat García Retortillo, and Silvia Minguez Maso. 2025. User Study Design for Identifying the Semantics of Bioethical Principles. In *Value Engineering in Artificial Intelligence*, Nardine Osman and Luc Steels (Eds.). Springer Nature, 22–39. https://doi.org/10.1007/978-3-031-85463-7_2

[45] Manel Rodríguez-Soto, Roxana Rădulescu, Filippo Bistaffa, Oriol Ricart, Arnau Mayoral-Macau, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Ann Nowé. 2026. Multi-objective reinforcement learning for provably incentivising alignment with value systems. *Artificial Intelligence* 351 (2026), 104460. https://doi.org/10.1016/j.artint.2025.104460

[46] Meg J Rohan. 2000. A Rose by Any Name? The Values Construct. *Personality and Social Psychology Review* 4 (2000), 255–277. Issue 3. https://doi.org/10.1207/S15327957PSPR0403_4

[47] Stuart Russell. 2022. Artificial Intelligence and the Problem of Control. In *Perspectives on Digital Humanism*, Hannes Werthner, Erich Prem, Edward A. Lee, and Carlo Ghezzi (Eds.). Springer, 19–24.

[48] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized Experience Replay. arXiv:1511.05952 [cs.LG] https://arxiv.org/abs/1511.05952

[49] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*. Vol. 25. Elsevier, 1–65.

[50] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 11.

[51] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael Wooldridge, Javier Morales, and Carlos Ansótegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1294–1302.

[52] Marc Serramia, Manel Rodríguez-Soto, Maite López-Sánchez, Juan Rodríguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansótegui. 2023. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines* 33 (11 2023), 761–790. https://doi.org/10.1007/s11023-023-09649-7

[53] Luciano C Siebert, Enrico Liscio, Pradeep K Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen Van Den Hoven, and Catholijn Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. *Frontiers in Artificial Intelligence and Applications* 354 (2022), 114 – 127. https://doi.org/10.3233/FAIA220193

[54] Walter Sinnott-Armstrong. 2023. Consequentialism. In *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[55] Benjamin J Smith, Robert Klassert, and Roland Pihlakas. 2022. Using soft maximin for risk averse multi-objective decision-making. *Autonomous Agents and Multi-Agent Systems* 37 (2022), 11. Issue 1. https://doi.org/10.1007/s10458-022-09586-2

[56] Nate Soares. 2018. *The Value Learning Problem*. Chapman and Hall/CRC, 89–97.

[57] Theodore Sumers, Robert Hawkins, Mark K Ho, Tom Griffiths, and Dylan Hadfield-Menell. 2022. How to talk so AI will learn: Instructions, descriptions, and autonomy. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22, Vol. 35)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.). Curran Associates, Inc., Red Hook, NY, USA, Article 2519, 14 pages. https://proceedings.neurips.cc/paper_files/paper/2022/file/e0cfde0ff720fa9674bb976e7f1b99d4-Paper-Conference.pdf

[58] Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. 2018. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology* 20, 1 (3 2018), 27–40. https://doi.org/10.1007/S10676-017-9440-6/FIGURES/1

[59] K. Van Moffaert, M.M. Drugan, and A. Nowé. 2013. Scalarized multi-objective reinforcement learning: novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, 191–199. https://doi.org/10.1109/ADPRL.2013.6615007

[60] Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press, Great Clarendon Street, Oxford, UK.

[61] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10582–10592. https://doi.org/10.18653/v1/2024.findings-emnlp.620

[62] Xinran Wang, Qi Le, Ammar Ahmed, Enmao Diao, Yi Zhou, Nathalie Baracaldo, Jie Ding, and Ali Anwar. 2025. MAP: Multi-Human-Value Alignment Palette. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. Curran Associates Inc., Red Hook, NY, USA, 81284–81313. https://openreview.net/forum?id=NN6QHwgRrQ

[63] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 2017. A Survey of Preference-Based Reinforcement Learning Methods. *Journal of Machine Learning Research* 18, 136 (2017), 1–46. http://jmlr.org/papers/v18/16-634.html

[64] C. F. Jeff Wu. 1983. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics* 11, 1 (1983), 95 – 103. https://doi.org/10.1214/aos/1176346060

[65] Andrea H. Wynn, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Learning human-like representations to enable learning human values. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 952, 31 pages.

[66] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, https://www.jmlr.org, Article 2322, 22 pages.

[67] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 1311, 12 pages. https://doi.org/10.5555/3454287.3455598

[68] Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. 2024. PROVABLE OFFLINE PREFERENCE-BASED REINFORCEMENT LEARNING. In *12th International Conference on Learning Representations, ICLR 2024*. Curran Associates Inc., Red Hook, NY, USA, 26112–26137. Publisher Copyright: © 2024 12th International Conference on Learning Representations, ICLR 2024. All rights reserved.; 12th International Conference on Learning Representations, ICLR 2024 ; Conference date: 07-05-2024 Through 11-05-2024.

[69] Huiying Zhong, Zhun Deng, Weijie J. Su, Zhiwei Steven Wu, and Linjun Zhang. 2024. Provable Multi-Party Reinforcement Learning with Diverse Human Feedback. arXiv:2403.05006 [cs.LG] https://arxiv.org/abs/2403.05006