

A Method for Learning Value Systems in Generative AI

Andrés Holgado Sánchez¹
Holger Billhardt^{1*}, Sascha Ossowski^{1*}

¹CETINIA, University Rey Juan Carlos
Unnumbered Tulipán Street, Móstoles, 28933 Madrid, Spain

Abstract

Value-aware AI systems require explicit computational representations of human values (*groundings*) and their aggregation into value systems in order to align their decisions with ours. As such representations are difficult to elicit, *value learning* seeks to infer them by observing human behaviour. This work addresses the lack of grounded value learning methods in generative AI: existing approaches typically replicate human preferences without awareness of the multidimensional structure of value alignment, or lack principled value system elicitation methods. To address these gaps, we adapt a previously validated *value system learning* method to the generative AI setting, which, based on pairwise prompt-response preference data, simultaneously learns: i) an implementation of a grounding for a set of values given by a multi-objective reward model, and ii) a value system representation in the form of a weighted linear scalarization of the previous grounding model. To ensure that the learned value systems are based on coherent value representations, our algorithm dynamically prioritizes the grounding learning process. We evaluate the method against baselines and a contemporary method on prompt-response preference datasets. Results show competitive performance and minimal trade-offs against the baselines, while improving explainability.

1 Introduction

Aligning AI systems with human intentions has long been recognized as a central challenge in artificial intelligence (Anderson, Anderson, and Armen 2006). With the widespread deployment of generative AI models – predominantly Large Language Models (LLMs) – the need for reliable mechanisms to steer their behaviour has become increasingly pressing. The dominant paradigm for this purpose is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017), which aligns a generative model with human preferences by (i) learning a reward model that estimates the *alignment* of its generations with human desires, and (ii) fine-tuning the generative model to maximize this reward using reinforcement learning.

A fundamental limitation of this approach is its inability to recognise the *pluralistic* nature of alignment (Sorensen

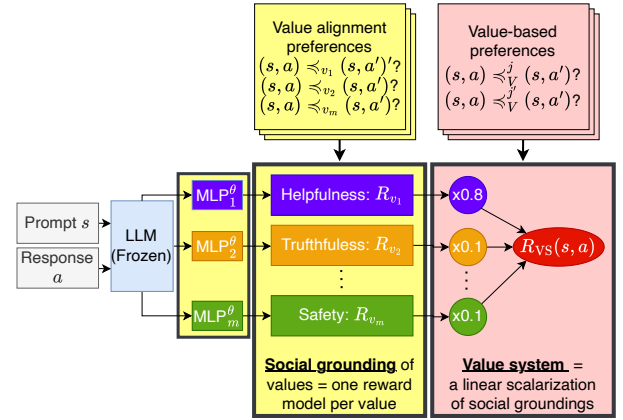


Figure 1: Diagram of the proposed value system learning solution. Based on pairwise preference data reflecting value alignment with multiple human values and the value-based preferences of multiple agents, we learn both a grounding for these values implemented by a multi-objective reward model and a linear value system representation in the form of linear scalarization weights.

et al. 2024), i.e. they do not decompose the *value alignment* (Russell 2022) of actions in terms of different *human values*. The field of *value awareness engineering* argues that AI systems should explicitly represent both values and the value-based preferences of agents (here called *value systems*) in order to be able to reason about them (Montes et al. 2023). With regard to values, the manual design of value representations is difficult and prone to misspecification (Sumers et al. 2022), especially if these representations are based on reward models (Hadfield-Menell et al. 2017). As an alternative, authors have proposed *value learning* (Soares 2018) as a means to acquire value representations –hereafter, *groundings* (Holgado-Sánchez et al. 2026)– automatically from stated preferences or demonstrations of behaviour. This idea has been explored in some decision-making contexts, where utility or reward functions represent morally relevant concepts and are subsequently used as de-

*These authors contributed equally.

cision objectives (Reddy et al. 2020; Rodríguez-Soto et al. 2026).

Following this idea, some works in the generative AI domain (e.g. (Dong et al. 2023; Wang et al. 2024a, 2025)) model alignment with multiple values using independent reward functions, employing a *multi-objective alignment* (Vamplew et al. 2018) framework. With regard to modelling the value systems of users, in those works it is assumed that users explicitly state their value systems directly (in the form of value weights or constraints). However, there is evidence that users are often unable to communicate their value system reliably (Siebert et al. 2022). Other methods approximate a Pareto frontier of aligned models for users to choose from (Zhou et al. 2024; Gupta et al. 2025), avoiding directly asking for the users’ value systems. However, this approach incurs in additional computational effort and user interactions. Finally, there are approaches that infer value systems from user data. Still, they tend to propose latent (Li et al. 2025) or opaque value system representations (Wang et al. 2024b) that have limited interpretability in terms of objectives/values.

To address these shortcomings, we propose a method for learning an approximation of a complete value system representation suited for preference modelling in generative AI applications, by means of a multi-objective decision making framework. Specifically, we learn a grounding of value labels implemented using reward models, and approximate value systems as weighted linear scalarizations over the previous value representations. Our algorithm jointly learns both representations from prompt–response preference pairs expressed in terms of the alignment of each response with each value, as well as of the alignment with the value system of a user, as illustrated in Figure 1. Notably, our optimization dynamically prioritizes the grounding learning process over the value system one, as careful value recognition is a precondition for value system estimation (Liscio et al. 2023).

This paper is organized as follows. In Section 2 we survey related works. In Section 3.2 we propose our value alignment and value system representation approach for generative AI models. In Section 4 we formalize our *value system learning* problem, and propose a theoretically sound algorithm that approximates a solution to it. Lastly, in Section 5, we evaluate our proposal and in Section 6 we provide conclusions, limitations and avenues for future work.

2 Related Works

2.1 Values in Generative AI

Recent work on AI alignment acknowledges that users demand different behavioural properties from generative AI assistants. Two principles arise as the most studied, namely *helpfulness* and *harmlessness/safety* (Bai et al. 2022b). Lately, other principles or goals such as *coherence* (Wang et al. 2024c), *objectivity* (Ji et al. 2024), *honesty* (Cui et al. 2023), or even *humor* (Köpf et al. 2023) have been explored and aligned for. These have often been considered as alignment targets or subgoals that describe different dimensions of the task of aligning assistant AI models with human values (Dognin et al. 2025; Kirk et al. 2024). These

alignment targets can be interpreted as domain-dependent “concepts” that realize broader human values (Osman and d’Inverno 2024). For instance, the *helpfulness* of a model may be considered a particularization of the more abstract value of *achievement* (Schwartz 1992), i.e. the value that fosters the pursuit of personal goals and aims.

2.2 Alignment With Given Value Systems

Authors typically introduce alignment targets/attributes/criteria (hereafter also referred to simply as *values*)¹ with the explicit goal of *steering* the behaviour of LLMs towards different value-based preferences, or *value systems* (Serramia et al. 2018), specified by the user. In this process, most methods learn or rely on an existing computable representation of the values, i.e. a grounding (Holgado-Sánchez et al. 2026), that measures the alignment or misalignment of responses with the values considered. This grounding notion is to some extent inherent in recent large language models (LLMs) and can be extracted using careful prompting strategies (Adams et al. 2025). However, for novel application contexts or previously unexplored values, it is generally necessary to train grounding models (normally using *reward models*) from human or AI annotations (Bai et al. 2022a). These value-alignment rewards can then be used to fine-tune a response generator (LLM-based), for instance through RLHF (Christiano et al. 2017).

Notable examples follow the previous paradigm. In (Dong et al. 2023), the authors learn a *multi-attribute reward model* that estimates the degree of alignment of prompt–response pairs with multiple values or attributes. They then fine-tune LLMs through supervised fine-tuning (SFT) to generate responses that adhere to user-specified desired alignment levels over these attributes. Similarly, (Wang et al. 2024a) learn a multi-objective reward model via linear regression on quantified value-annotated examples, and subsequently train an LLM to combine these value-specific rewards according to given reward weightings. Other approaches aim to control the trade-offs between alignment with different values, such as Controllable DPO (Guo et al. 2024) and MAP (Wang et al. 2025), by enforcing manually specified value-alignment constraints. Finally, some authors focus on obtaining a Pareto front of LLMs aligned for multiple value combinations (Yang et al. 2024; Gupta et al. 2025). These approaches, however, may expend computational resources on approaching regions of the Pareto front that are ultimately undesirable for the user.

2.3 Estimating Value Systems

The previous works proposed value-steerable models that learn value representations (groundings), however, to adequately steer their models according to particular users, they expect them to convey their preferences in the form required by their respective architectures (e.g. weights). How-

¹To simplify the reading, we will refer to all alignment targets or related goals (such as *helpfulness*) as full-fledged “values”, assuming they have already been identified as relevant alignment targets in the application context considered; see value identification (Liscio et al. 2021).

ever, eliciting the value system of a user in terms of context-dependent values, is, in general, a hard problem that has received the name of *value system estimation* (Siebert et al. 2022; Liscio et al. 2023). Classical solutions for this problem include surveys (Schwartz 1992) and participatory design methodologies (Ziegfeld et al. 2025). However, it has been shown that value systems elicited in surveys may change depending on the task the user is trying to perform (Kirk et al. 2024), which makes translating these value system estimations into actionable model behaviour a challenging task.

Therefore, authors tend to learn value system representations directly from application-oriented data. For example, if a set of models aligned with known value system representations is available, selecting an appropriate model by asking an user to judge responses generated by each model is a natural approach to indirectly estimate her value system (Ramé et al. 2023; Gupta et al. 2025), albeit a costly one. A more scalable solution is proposed by Wang et al. (Wang et al. 2024b), who learn value-specific reward models and infer prompt-dependent value system representations in the form of scalarization weights via a dedicated deep neural network. While effective, this design lacks interpretability, as the relationship between prompt features and the inferred value system weights is not directly traceable. Relatedly, works such as LoRe (Bose et al. 2025) and PAL (Chen et al. 2025) learn latent preference embeddings to efficiently capture personal user preferences (i.e. value systems), but the dimensions of these embeddings do not directly correspond to identifiable values. Unlike these approaches, our method aims to learn a transparent value system representation based on represented values, which allows an agent to reason explicitly on the values and how they combine into a value system.

3 Value Alignment in Generative AI

Following common practice, we model the generative task of any given generative model (e.g. LLM) as a contextual bandit problem (Tennant, Hailes, and Musolesi 2025) (or MDP, in general), where there is a set of states or instructions (from those of a possibly infinite set S , e.g. prompts) for which different actions (from a set A , e.g. responses) can be generated or selected. The model is described by a policy $\pi(a | s)$ that outputs the most appropriate action $a \in A$ given a state $s \in S$. We define the pair $\tau = (s, a)$ as a *generation* and the set of generations as $\mathcal{T} \subset S \times A$.

We assume that appropriateness of a generation $\tau = (s, a)$ can be modelled through a quantitative reward function (or *reward model* in the generative AI jargon) $R(s, a)$. Such reward model can be learned from datasets of human annotated examples of preferences (Christiano et al. 2017) or, directly, from scored outputs to example user inputs (Cui et al. 2023). The goal of a model π that is aware of this reward model is to generate actions that maximize the expected reward.

In the following, we will be interested in measuring and learning the appropriateness of generations solely in terms of their alignment with human values and, based on that, their appropriateness in terms of the value systems of particular agents, using several reward models. In this paper, we leave out of scope the fine-tuning process of policies with

these reward models, that may be carried out using existing methods such as RLHF (Christiano et al. 2017; Bai et al. 2022b).

3.1 Representing Value Alignment

To properly introduce our reward learning objectives, we will adopt the theoretical value alignment representation assumptions previously proposed in (Holgado-Sánchez et al. 2025) and adapt these to the generative AI context.

Definition 1 (Value Alignment) *The alignment of a set of generations $\mathcal{T}$ with a value v_i (in general, the alignment preferences with v_i) is represented by a weak order $\preceq_{v_i}$ over $\mathcal{T}$, where $\tau \preceq_{v_i} \tau'$ means that the generation τ' is at least as aligned with value v_i as τ .*

Then, to jointly specify the alignment with a set of values, we formalize the notion of *grounding*.

Definition 2 (Grounding) *A grounding of the set of values V is a set of weak orders $\preceq_V = \{\preceq_{v_i}\}_{i=1}^m$.*

To facilitate the computational representation of these value alignment preferences, we propose approximating them using separate utility models for different values. This view has been paramount in works that model human/moral values in general (Rodríguez-Soto et al. 2022; Serramia et al. 2018; Montes and Sierra 2022; Karanik et al. 2024). Thus, we propose a *value alignment function*, $\mathcal{A}_{v_i}$, that quantifies the value alignment of a generation with a value $v_i \in V$, i.e. that represents the relation $\preceq_{v_i}$: i.e., for all $\tau, \tau' \in \mathcal{T}$: $\tau \preceq_{v_i} \tau' \iff \mathcal{A}_{v_i}(\tau) \leq \mathcal{A}_{v_i}(\tau')$.

In the generative AI setting, an alignment function can be directly implemented through a reward model: $R_{v_i}(\tau) \equiv \mathcal{A}_{v_i}(\tau)$. Thus, gathering alignment functions for each preference relation, we can represent a grounding $\preceq_V$ with a multi-objective reward model, or *reward vector* $\mathbf{R}(s, a)$ such that $\mathbf{R}(\tau) = (R_{v_1}(\tau), \dots, R_{v_m}(\tau)) \equiv (\mathcal{A}_{v_1}(\tau), \dots, \mathcal{A}_{v_m}(\tau))$. Given a reward vector, we define the particular grounding that it implements as $\preceq_{\mathbf{R}} = \{\preceq_{R_{v_i}}\}_{i=1}^m$, i.e. $\tau \preceq_{R_{v_i}} \tau' \iff R_{v_i}(\tau) \leq R_{v_i}(\tau')$. In occasions, we will refer to this reward vector also as the *grounding function/model* to emphasize its role as the implementation of a grounding.

Importantly, our reward functions do not aim to capture the whole meaning of the given values, rather, we are estimating the *alignment* of actions with these value based on the proposed multi-objective model.

3.2 Representing Value Systems

A *value system* expresses the general importance assigned to each value by a certain person or group (in general, an “agent”) in a specific context. Given a grounding, it represents the individual (value-based) preferences over generations of an agent.

Definition 3 (Value system) *The value system of an agent j , based on a grounding $\preceq_V$, is determined by a weak order $\preceq_V^j$ over $\mathcal{T}$. If $\tau \preceq_V^j \tau'$, we say that τ is equally or more aligned than τ' with j ’s value system.*

The idea is that the value system preferences of an agent are based on the individual value preferences. Following other works in the field (Serramia et al. 2018; Rodríguez-Soto et al. 2026), we assume that the value system of an agent can be approximated by a linear combination of the alignment of a generation with the values. Thus, our value system model is *welfarist utilitarian* (Sinnott-Armstrong 2023), i.e. values are conceived as different sources of good, and each agent’s value system weighs their relative importance.

Definition 4 (Value System Function) Let j be an agent with a value system $\preceq_V^j$ and a grounding represented by a reward vector $\mathbf{R}$. The function $\mathcal{A}_{W_j, \mathbf{R}}(\tau) = W_j \cdot (R_{v_1}(\tau), \dots, R_{v_m}(\tau))^T$ is a **value system function** for j if it represents $\preceq_V^j$ over $\mathcal{T}$, i.e., for all $\tau, \tau' \in \mathcal{T}$:

$$\mathcal{A}_{W_j, \mathbf{R}}(\tau) \leq \mathcal{A}_{W_j, \mathbf{R}}(\tau') \iff \tau \preceq_V^j \tau'$$

where $W_j = (w_j^{v_1}, \dots, w_j^{v_m})$ are called *value system weights*. These are bounded in the unit $(m-1)$ -simplex²: $W_j \in \Delta^{m-1}$, i.e. $W_j \in (0, 1)^m$, $\sum_{i=1}^m w_j^{v_i} = 1$.

We can interpret the previous definition as having a *value system reward* for each agent that is a linear scalarization (Van Moffaert, Drugan, and Nowé 2013) of the multi-objective reward vector $\mathbf{R}$ with weights W_j :

$$R_j(\tau) = W_j \cdot \mathbf{R}(\tau)^T$$

Finally, we denote the value system implemented through the value system alignment function $\mathcal{A}_{W_j, \mathbf{R}}$ (with weights W_j and reward vector $\mathbf{R}$) by $\preceq_{\mathbf{R}}^{W_j}$, i.e. $\tau \preceq_{\mathbf{R}}^{W_j} \tau' \iff W_j \cdot \mathbf{R}(\tau) \leq W_j \cdot \mathbf{R}(\tau')$.

We emphasize that our value system representation is a mere approximation of the true value systems of agents, which might be much more complex and context-dependent. We chose a linear representation that eases the interpretability of our model.

4 Learning Value Systems

In this section, we present our proposal: a model and algorithm for learning simultaneously a reward vector $\mathbf{R}$ and a weight vector W , which approximate a grounding of the values considered and the value system of a certain agent, respectively. Our learning approach relies on a dataset of annotated preferences for both, values and the value system of an agent.

²As we treat values as sources of good, we do not consider value systems that willingly demote values by using negative weights. If demoting a value is expected to occur, the designer should instead include another value in the model as the negated version of the original. An example of such case is the technical value of response *complexity* (Wang et al. 2024c), which in some situations might be undesirable (when a response is for a audience of children, for instance). Instead of using a negative value system weight, we suggest adding the contrary value of *simplicity*, to represent the abstract goal of reducing complexity.

4.1 Data Assumptions and Scope

Within the field of generative AI, even though datasets that identify individual annotators or agents exist (Köpf et al. 2023; Kirk et al. 2024), the majority of datasets that label examples along multiple alignment dimensions or values agglomerate the opinions of distinct individuals with potentially divergent preferences (Wang et al. 2024c; Ji et al. 2023, 2024). Our work seeks to be applicable in the latter cases as well, assuming that the dataset preferences are those of an hypothetical abstract agent representing the collective.

This assumption may raise concerns about the usefulness of the proposed approach, as it might attempt to reconcile potentially divergent opinions into a single representation. However, with regard to value recognition in particular, we argue this view might not be problematic. Namely, although we consider agents may hold divergent interpretations of the values under consideration (for example, the meaning of “freedom”), within delimited domains we can reasonably assume that most agents understand values in broadly similar manner. That is, there exists a degree of consensus regarding the meaning of values, which we refer to as a social grounding (Holgado-Sánchez et al. 2025).

In the generative AI domain, the existence of such social groundings is often implicitly assumed. We identify two main reasons for this. First, most alignment-related datasets justify the inclusion of multiple alignment criteria as universal quality dimensions and provide explicit definitions for each, which condition the preference annotation process (Bai et al. 2022b). Second, some works explicitly average different opinions over the same utterances to improve “robustness” of the annotations (Wang et al. 2024c), to then train a shared “grounding-function-like” model.

With regard to value systems, it is clear that even in the same application domain, agents may have diverse value-based preferences (value systems). Our work, however, performs an implicit aggregation (Lera-Leri et al. 2024) of potentially divergent value systems. Doing this has been shown to have limitations in terms of the final representativeness of value-based preferences, specially in datasets with heterogeneous agents (Holgado-Sánchez et al. 2025; Chakraborty et al. 2024). We acknowledge potential limitations of the representativeness that we achieve in these cases, but argue that our approach is still valuable, as the learned value system can be interpreted as a transparent representation of the overall, empirical, importance given to each value according to the dataset annotators.

In conclusion, in this paper, we consider that the value alignment preferences observed in our dataset correspond to a social grounding in a delimited context (e.g. text generation) and the value system preferences pertain to some abstract agent. When data annotations originate from a group of agents with potentially heterogeneous value systems, then our learning approach would try to identify the aggregated value system that best represents the collective preferences of that group.

Importantly, due to the limitations mentioned above, if our model is learned from heterogeneous datasets, the resulting value system reward (given by $W \cdot \mathbf{R}$) should not be intended for use as a stand-alone objective in RLHF

fine-tuning. In such cases, it would likely bias the generative model toward majority opinions. Instead, we recommend using the learned grounding function $\mathbf{R}$ for multi-objective fine-tuning. Specifically, generative models should be trained in a way that allows their immediate alignment with respect diverse value systems, following approaches similar to prior work (Gupta et al. 2025; Wang et al. 2024c; Jang et al. 2024). Our approach offers an additional advantage: it enables the exploration of value systems in the vicinity of the learned one, that acts as the representation of the preferences of the majority.

4.2 Problem Definition

Considering the above, our *value system learning* problem consists of *learning* a grounding implementation together with a representation of the value system of a certain (possibly abstract) agent j that best approximates the social grounding $\preceq_V$ and value system $\preceq_V^j$ enacted in a observed preference dataset. In particular, we assume access to examples of value alignment and value system preferences over pairs of generations. Formally, we assume a dataset, DS, composed by entries of the type $(\tau, \tau', y_V^j, y_{v_1}, \dots, y_{v_m})$, where each *label* $y_- \in \{0, 0.5, 1\}$ indicates the preference of τ over τ' (1), the contrary case (0) or indifference (0.5). Specifically, label y_V^j indicates the preference according to the agent’s value system, and labels $y_{v_i}, i = 1, \dots, m$ represent the preference regarding the alignment of each generation pair with respect to each value v_i . As mentioned above, we assume that the latter labels correspond to the social grounding of the values.

Given the analysis in Section 3.2, we propose reducing our learning problem into two tasks i) learn a quantitative reward vector $\mathbf{R}$ such that the grounding $\preceq_R$ implemented by $\mathbf{R}$ approximates the real social grounding $\preceq_V$ and ii) learn a vector of value system weights W such that the value system $\preceq_R^W$ approximates the real value system $\preceq_V^j$.

We now provide ways to quantitatively *measure* the appropriateness of potential candidate solutions. For that, we quantify the alignment preference differences between our models and the given data. To measure preference differences in general, we propose the *discordance* between preference relations ($\preceq^1, \preceq^2$) as the proportion of ordered pairs of generations in a set $S \subseteq \mathcal{T} \times \mathcal{T}$ that are ranked differently:

$$d_S(\preceq^1, \preceq^2) = \frac{1}{|S|} \sum_{(\tau, \tau') \in S} \mathbb{1}((\tau \preceq^1 \tau') \neq (\tau \preceq^2 \tau')) \quad (1)$$

In our settings, the aim is to learn grounding functions and value system weights with minimal discordance to the original data (represented in the dataset DS). Thus, we can define our value system learning problem as a nested optimization problem:

$$\begin{aligned} (W^*, \mathbf{R}^*) \in & \arg \min_{W, \mathbf{R}' = (R_{v_1}, \dots, R_{v_m})} d_{DS}(\preceq_{\mathbf{R}'}^W, \preceq_V^j) \\ \text{subject to } & R'_{v_i} \in \arg \min_{R_{v_i}} d_{DS}(\preceq_{R_{v_i}}, \preceq_{v_i}), \\ & \forall i \in \{1, \dots, m\} \end{aligned} \quad (2)$$

The formulation prioritizes the minimization of the average discordance on the learned grounding functions, before improving the value system estimation. This ensures that the learned value system weights are based on a coherent social grounding approximation.

It should be noted that in Problem 6, the discordance can be directly calculated by comparing the preference predictions of the learned models with the original preference relations on the generation pairs included in the dataset DS. To do so, we discretize the score differences predicted by our model for each pair of generations to predict strict preference ($\tau \prec \tau' / \tau \succ \tau'$) or indifference ($\tau \simeq \tau'$), which is treated as predicting both $\tau' \preceq \tau$ and $\tau' \succeq \tau$. Then, we check the agreement of the discretized predictions with the corresponding dataset label $y \in \{0, 1, 0.5\}$. Details on how the predictions of the model are discretized and the data-based discordance is calculated are given in supplementary material (Holgado-Sánchez, Billhardt, and Ossowski 2026).

4.3 Algorithm

Our algorithm is an approximate solution to Problem 2 that versions the classical deep reward learning method based on the Bradley-Terry (BT) model (Bradley and Terry 1952). The solution is represented by several neural networks. First, a reward vector network $\mathbf{R}^\theta$ with parameters θ implements an approximation the social grounding in the data. Second, a linear layer given by certain value system weights W^ω that are parametrized with $\omega \in \mathbb{R}^m$ through a *softmax* calculation: $W^\omega = (w^{v_1}, \dots, w^{v_m}) = \frac{\exp \omega}{\sum \exp(\omega)}$, so W^ω remains in the simplex.

For learning purposes, we approximate the value alignment and value system preferences in a differentiable manner with the BT model. Namely, we use the sigmoid (σ) of the reward difference between two generations that share the same initial prompt: $p(\tau \succ \tau' \mid R) = \sigma(R(\tau) - R(\tau'))$, as an approximation of the degree of preference (given as a probability between 0 and 1) of preferring the generation τ over τ' . To learn with this model, we modify the parameters of R so that the values of $p(\tau \succ \tau' \mid R)$ are closer to the preference labels $y \in \{0, 0.5, 1\}$ by minimizing a binary cross-entropy loss (Christiano et al. 2017). Minimizing this loss, in Eq. 3, discordance between the model R and the dataset will also be decreased.

We add a modification to the loss using the centering mechanism proposed in (Eisenstein et al. 2024), which reduces the magnitude of the estimated rewards and helps ensure that reward scales are comparable across values. This mechanism introduces a hyperparameter r , which we set to the recommended value of $r = 0.01$.

$$\begin{aligned}\mathcal{L}(\tau, \tau', y \mid R) = & -y \log p(\tau \succ \tau' \mid R) \\ & - (1 - y) \log (p(\tau \prec \tau' \mid R)) \\ & + r (R(\tau) + R(\tau'))^2\end{aligned}\quad (3)$$

Then, we propose a loss associated to the discordance of our model with the alignment preference examples for each value $v_i \in V$ in our dataset (Eq. 4).

$$\mathcal{L}_{v_i}^\theta(\text{DS}) = \sum_{(\tau, \tau', y_{v_i}) \in \text{DS}} \frac{\mathcal{L}(\tau, \tau', y_{v_i} \mid \mathbf{R}^\theta)}{|\text{DS}|} \quad (4)$$

We consider all the previous losses together to define a *grounding (discordance) loss* given by: $\mathcal{L}_V^\theta(\text{DS}) = (\mathcal{L}_{v_1}^\theta(\text{DS}), \dots, \mathcal{L}_{v_m}^\theta(\text{DS}))$.

And we define a loss associated to value system discordance (Eq. 5).

$$\mathcal{L}_{\text{VS}}^{\omega, \theta}(\text{DS}) = \sum_{(\tau, \tau', y_V^j) \in \text{DS}} \frac{\mathcal{L}(\tau, \tau', y_V^j \mid W^\omega \cdot \mathbf{R}^\theta)}{|\text{DS}|} \quad (5)$$

A naïve approach to use these losses to solve Problem 2 would be first minimizing the average of the grounding discordance losses $\mathcal{L}_V^\theta$, and then, with θ frozen, learn the weights that minimize the value system discordance loss $\mathcal{L}_{\text{VS}}^{\omega, \theta}$. However, it has been shown that multiple grounding functions are compatible with a low discordance while only a subset of those can also be effectively combined linearly to approximate value system (Holgado-Sánchez et al. 2025).

Therefore, we propose a *simultaneous* learning approach that balances both goals. The idea is treating Problem 2 as a constrained optimization problem, where the constraint is minimizing the inner grounding discordance objective. In this formulation (Problem 6), we introduce the constraints that the grounding discordance losses in Eq. 4 should reach an (unknown, approximated) per-value *loss target* $\mathcal{L}_{v_i}^*$ that depends on the capabilities of the available model. These targets $\{\mathcal{L}_{v_i}^*\}_{i=1}^m$ are estimated during training by an “exponentially weighted *minimum*” of observed losses during training (see step 3 of the **Algorithm description**).

$$\begin{aligned}(\omega^*, \theta^*) \in & \arg \min_{W, \mathbf{R}'} \mathcal{L}_{\text{VS}}^{\omega, \theta'}(\text{DS}) \\ \text{subject to } & \theta' \in \{\theta \mid \mathcal{L}_{v_i}^\theta(\text{DS}) \leq \mathcal{L}_{v_i}^*\} \\ & \forall i \in \{1, \dots, m\}\end{aligned}\quad (6)$$

To approach Problem 6, we obtain its Lagrangian (Eq. 7). It introduces Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^+$. Then, the optimization consists of finding a saddle point for the *min-max* goal in Eq. 8.

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\lambda}}^{\omega, \theta}(\text{DS}) = & \mathcal{L}_{\text{VS}}^{\omega, \theta}(\text{DS}) \\ & + \sum_{i=1}^m \lambda_i (\mathcal{L}_{v_i}^\theta(\text{DS}) - \mathcal{L}_{v_i}^*)\end{aligned}\quad (7)$$

$$\min_{\theta, \omega} \left[\max_{\boldsymbol{\lambda} > 0} [\mathcal{L}_{\boldsymbol{\lambda}}^{\omega, \theta}(\text{DS})] \right] \quad (8)$$

Assuming that the reward vector $\mathbf{R}^\theta$ is an affine transformation (e.g. obtained by training only an additional single-layer perceptron over embeddings extracted by a “base” LLM), Problem 6 is convex, and with additional minor assumptions, strong duality holds (Boyd and Vandenberghe 2004), which justifies the theoretical use of dual ascent (Boyd et al. 2011) to find an optimal solution to Problem 8 (for fixed loss targets). Instead of dual ascent, which requires the exact minimization of the Lagrangian at every training iteration with respect to ω and θ , we approximate a solution with a modified gradient descent-ascent (GDA) algorithm (Zamani, Abbaszadehpour, and de Klerk 2022), which converges to a local Nash equilibrium and even the global optimum with certain learning rate choices (Fallah, Ozdaglar, and Pattathil 2020). See supplementary material (Holgado-Sánchez, Billhardt, and Ossowski 2026), Section A.2 for proofs of the previous theoretical aspects.

Before introducing the algorithm, we add some technical modifications. First, to guarantee that the learning step sizes on the Lagrangian remain constant, we add an auxiliary “value system multiplier”, λ_{VS} , which is calculated so that the sum of the multipliers including this one have an average value of 1³. Second, to avoid any multiplier (including λ_{VS}) from vanishing, we force them to be over a certain user-defined threshold $\lambda_0 > 0$. To satisfy the previous constraints, we use auxiliary parameters $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_1, \dots, \bar{\lambda}_m, \bar{\lambda}_{\text{VS}}) \in \mathbb{R}^{m+1}$ to parametrize the “grounding multipliers” $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ and the value system one λ_{VS} , using: $(\lambda_1, \dots, \lambda_m, \lambda_{\text{VS}}) = \frac{(m+1) \exp \bar{\boldsymbol{\lambda}}}{\sum_{\bar{\lambda} \in \bar{\boldsymbol{\lambda}}} \exp \bar{\boldsymbol{\lambda}}} (1 - \lambda_0) + \lambda_0$. Finally, for theoretical reasons, we add ℓ_2 regularization to the Lagrangian, affecting all parameters $\omega, \theta, \bar{\boldsymbol{\lambda}}$. The Lagrangian used in our algorithm with the previous modifications is in Eq. 9.

$$\begin{aligned}\bar{\mathcal{L}}_{\bar{\boldsymbol{\lambda}}}^{\omega, \theta}(\text{DS}) = & \lambda_{\text{VS}} \mathcal{L}_{\text{VS}}^{\omega, \theta}(\text{DS}) + \gamma_\omega \|\omega\|^2 + \gamma_\lambda \|\bar{\boldsymbol{\lambda}}\|^2 \\ & + \sum_{i=1}^m \lambda_i (\mathcal{L}_{v_i}^\theta(\text{DS}) - \mathcal{L}_{v_i}^*) + \gamma_\theta \|\theta\|^2\end{aligned}\quad (9)$$

Algorithm description. After random parameter initialization, our algorithm proceeds by repeatedly executing the following three steps over all batches in the dataset, for N epochs⁴:

1. **Gradient descent.** Calculate gradients of Eq. 9 with respect to the model parameters (ω and θ) for several batches (accumulating them). Then, perform gradient *descent* step to update ω and θ .
2. **Gradient ascent.** We perform a gradient *ascent* step on the Lagrange multiplier parameters $\bar{\boldsymbol{\lambda}}$ based on the Lagrangian. In practice, for this update, we compute the gradient of a modified version of the Lagrangian in Eq. 7

³A similar multiplier normalization process was proposed for Safe RLHF methods (Dai et al. 2024).

⁴An epoch here is a pass over the complete dataset.

with grounding discordance losses measured on previous batches (BS):

$$\nabla_{\bar{\lambda}} \left[\sum_{i=1}^m \lambda_i \max(\mathcal{L}_{v_i}^\theta(\text{BS}) - \mathcal{L}_{v_i}^*, 0) + \gamma_\lambda \|\bar{\lambda}\|^2 \right]$$

With the “max” projection, this gradient is approximately zero (up to the effect of ℓ_2 regularization) for those multipliers whose associated grounding discordance losses, computed over recent batches, have already met their target values. This design ensures that the primary objective remains reducing grounding discordance losses as much as possible, even beyond the target estimations. The multipliers can still decrease slightly due to the ℓ_2 regularization term, which helps correct potential overestimations. It is noticeable that we do not update the value system multiplier λ_{VS} based on the value system discordance loss. However, when the gradients of the grounding multipliers ($\lambda = (\lambda_1, \dots, \lambda_m)$) vanish, the ℓ_2 regularization progressively reduces the gap between λ_{VS} and λ . Thus, only once the grounding discordance losses have reached their targets, the relative priority of minimizing the value system discordance loss is subject to a slight increase.

3. **Loss target update.** After repeating steps 1-2 for a number of steps and batches (let this number be $u > 0$), we re-estimate the grounding discordance loss targets that are attainable with our reward model R^θ . To do so, we employ a “exponentially weighted *minimum*” with coefficient $\eta \in (0, 1)$ (close to 1) of the averages of the losses obtained during each period of u steps. Namely, given the calculation of the loss associated with each value v_i across u different batches $\{\text{BS}_k\}_{k=1}^u$, and its current loss target estimation $\mathcal{L}_{v_i}^*$, the update is: $\mathcal{L}_{v_i}^* \leftarrow \eta \mathcal{L}_{v_i}^* + (1 - \eta) \min\{\mathcal{L}_{v_i}^*, \frac{1}{u} \sum_{k=1}^u \mathcal{L}_{v_i}^\theta(\text{BS}_k)\}$. We set the initial value of the target losses as the average of the losses obtained in the first u batches.

We supply a fully-detailed algorithm pseudocode in the supplementary material (Holgado-Sánchez, Billhardt, and Ossowski 2026).

5 Evaluation

We evaluate our proposal by comparing its performance against two baselines and the Armo-RM method (Wang et al. 2024b), a competitive reward model with the most similar architecture to our approach we found in the literature. We analyse 2 datasets: ULTRAFEEDBACK (Cui et al. 2023) and the text-to-text task in PKU-ALIGN-ANYTHING (Ji et al. 2024). We chose these datasets since they have been widely used for LLM alignment and because their structure is close to our dataset assumptions: indeed, in PKU-ALIGN-ANYTHING the annotators are not identified and in ULTRAFEEDBACK, the annotator is a single GPT-4 agent.

Methods. Here we detail each method characteristics and training setting:

- **Armo-RM.** A pretrained multi-objective reward model that predicts alignment with 19 *attributes* (that we treat

as values) from six datasets. It uses a frozen *Llama3-8B-v0.1* backbone and a linear head over the last hidden state acting similar to our reward vector R (grounding function). A scalar reward is obtained by linearly combining these outputs with weights $W_{\text{arm}}(p) \in [-1, 1]^{19}$, produced by a prompt-conditioned three-layer MLP trained with MSE.

- **VSL-RM (ours).** We use the same reward vector architecture as Armo-RM (a linear layer over the frozen *Llama3-8B-v0.1* hidden state). We train it with our algorithm from scratch on both datasets.
- **BT-RM.** A single-objective model that minimizes only the value system discordance loss (Eq. 5), without observing grounding discordance losses. It shares the same architecture and parameters as VSL-RM, and we also train one version per dataset. It serves as a lower bound on attainable value system discordance with the used architecture.
- **SEQ-RM.** Shares the same architecture as VSL-RM but is trained sequentially: first to solely minimize grounding discordance (updating θ), then minimizing the value system discordance (updating ω with θ frozen) for an additional 50% of epochs and higher learning rate. It serves both as a control baseline for comparing sequential vs. simultaneous training and as oracle for minimum attainable grounding discordance with the used architecture.

Training details. The datasets are split into training, validation, and test sets. The validation data is used for hyperparameter selection, while test data is used only for final evaluation. All models are trained with four random seeds (shared across methods), except Armo-RM, which is used as provided on HuggingFace⁵.

We trained the models (except Armo-RM) using approximately the same set of common hyperparameters selected with a standard search method. The baseline models rely primarily on standard hyperparameters (e.g., batch size), which have limited impact on performance in this context, so we keep them aligned with those used for VSL-RM without further tuning. The only exception is SEQ-RM, for which we increase the value system learning rate to compensate for the smaller number of value system training iterations.

All experiments were repeated with four random seeds. We observed little variation across runs, and therefore decided not to increase the number of training seeds. As optimizer, we used a paged, 32 bit “AdamW” one. Further details on hyperparameters, dataset preprocessing, and hardware are provided in supplementary material (Holgado-Sánchez, Billhardt, and Ossowski 2026), Section A.4.

Metrics. For an easier interpretation of results, we report grounding and value system *accuracies*, which we define as the complement of the grounding and value system discordances, respectively:

$$\text{acc}_{\text{DS}}(R_{v_i}, \preceq_{v_i}) = 1 - d_{\text{DS}}(\preceq_{R_{v_i}}, \preceq_{v_i}),$$

$$\text{acc}_{\text{DS}}((W, R), \preceq_V^j) = 1 - d_{\text{DS}}(\preceq_R^W, \preceq_V^j).$$

⁵Armo RM model: <https://huggingface.co/RLHFlow/ArmoRM-Llama3-8B-v0.1>, version from May 11th, 2026.

Accuracy tends to 1 when each model perfectly reproduces the preferences of the data (corresponding to a discordance of 0), and tends to 0 when discordance raises to its maximum (1).

5.1 Results

Figure 2 presents training curves for the *average grounding accuracy* (denoted by $\text{AGA} = \sum_{i=1}^m \frac{1}{m} \text{acc}_{\text{DS}}(Rv_i, \prec_{v_i})$) and the *value system accuracy* (denoted by $\text{VSA} = \text{acc}_{\text{DS}}((W, \mathbf{R}), \prec_V^j)$). The curves depict the evolution of these accuracies throughout training, measured on the validation splits of each dataset. We do not present the curves for Armo-RM, as it was originally pretrained.

As expected, BT-RM achieves highest VSA, as it is optimized solely for this objective, while VSL-RM clearly outperforms it in AGA due to the inclusion of the grounding discordance loss. Interestingly, AGA increases rapidly in both methods. Likely, this occurs because of our architecture where the value system reward is a positive combination of a reward vector (corresponding to the different values). Thus, a better prediction of the preferences of the value system correlates with a better value alignment estimation. Also in the case of BT-RM a reward vector is learned. However, this vector would not directly correspond to the underlying values. Still, the learned vector is indirectly correlated with the original values. This explains the fact that AGA increases rapidly for BT-RM (due to the positive correlations) and then drops (since the reward vector does not actually represent the original values). In VSL-RM, AGA still improves over the learning epochs.

For SEQ-RM, both AGA and VSA improve during the first phase in both datasets, with VSA computed using uniform value system weights ($1/m$)⁶. Despite its simultaneous learning approach, VSL-RM matches the AGA achieved by SEQ-RM, which is optimized solely for grounding discordance. This indicates that VSL-RM effectively prioritizes learning grounding functions with the lowest discordance, as intended in Problem 2. In the second phase of SEQ-RM, VSA increases further, almost matching the one achieved with VSL-RM. We will show the gap is still statistically significant, by evaluating the test datasets.

Table 1 (for ULTRAFeEDBACK) and Table 2 (for PKU-ALIGN-ANYTHING) present the performance of the models on the **test datasets**. We omit the presentation of grounding accuracies in the case of BT-RM, as the model predictions in value alignment are latent constructs used to minimize value system discordance.

All methods achieve comparable VSA across datasets, with BT-RM consistently performing best and Armo-RM exhibiting the weakest results. In the PKU-ALIGN-ANYTHING dataset, this could be expected, because Armo-RM was not trained on it. However, Armo-RM was trained on ULTRAFeEDBACK, and we expected that Armo-RM’s gating mechanism should, in principle, outperform our lin-

⁶VSA increases here due to the architectural effect described before: a good estimation of value alignment preferences, combined with positive weights, leads to a better estimation of value-based preferences.

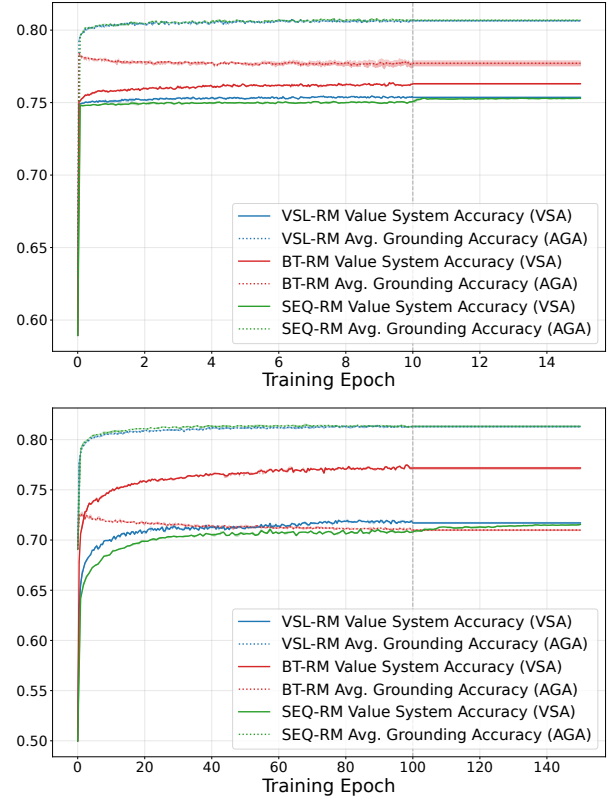


Figure 2: Average grounding and value system accuracy per training method (VSL-RM, BT-RM, SEQ-RM) in the validation splits of ULTRAFeEDBACK (Top) and PKU-ALIGN-ANYTHING (Bottom). It might be difficult to appreciate, but AGA by VSL-RM and SEQ-RM almost coincide. The vertical line indicates the training moment where VSL-RM and BT-RM end their training and SEQ-RM starts its second phase to maximize VSA.

ear value system representation in predicting overall preferences. This inaccuracy might be due to the fact that its training shared focus with other 5 datasets with possibly divergent overall preference criteria. As observed in Figure 2, BT-RM slightly outperforms VSL-RM regarding VSA in ULTRAFeEDBACK and more substantially in PKU-ALIGN-ANYTHING, suggesting that, in the latter, value systems may not easily be approximated as linear functions of the dataset value alignment scores. This indicates that more expressive value system models will be necessary to improve performance.

Comparing VSL-RM with SEQ-RM, the former consistently achieves marginally higher VSA, yet with statistically significant differences (one-sided Welch t-tests: $p = 0.0004$ in PKU-ALIGN-ANYTHING and $p < 0.0001$ in ULTRAFeEDBACK). This highlights a limitation of sequential optimization: although multiple grounding functions can achieve similar AGA, some are more suitable for linear aggregation when minimizing value system discordance, as have been shown (Holgado-Sánchez et al. 2025). The rel-

Method	Helpfulness	Honesty	Truthfulness	Instruction Following	AGA	VSA
Armo-RM	0.826	0.814	0.840	0.824	0.826	0.754
BT-RM	—	—	—	—	—	0.768 ±0.0002
VSL-RM	<u>0.812</u> ±0.0005	<u>0.793</u> ±0.0002	0.786 ±0.0007	<u>0.815</u> ±0.0003	0.802 ±0.0002	<u>0.760</u> ±0.0003
SEQ-RM	<u>0.813</u> ±0.0003	<u>0.793</u> ±0.0003	<u>0.788</u> ±0.0008	<u>0.815</u> ±0.0003	<u>0.802</u> ±0.0002	0.758 ±0.0002

Table 1: Average grounding accuracy and value system accuracy (ULTRAFEEDBACK).

Method	Prompt Following	Objectivity	Clarity	Information Richness	Safety	AGA	VSA
Armo-RM	0.715	<u>0.798</u>	0.609	0.647	0.837	0.721	0.685
BT-RM	—	—	—	—	—	—	0.773 ±0.0008
VSL-RM	0.765 ±0.0029	<u>0.798</u> ±0.0030	0.720 ±0.0022	0.819 ±0.0029	<u>0.890</u> ±0.0024	<u>0.799</u> ±0.0005	<u>0.714</u> ±0.0009
SEQ-RM	0.762 ±0.0014	0.806 ±0.0031	0.721 ±0.0024	0.818 ±0.0042	0.896 ±0.0014	0.801 ±0.0015	0.707 ±0.0015

Table 2: Average grounding and value system accuracies (PKU-ALIGN-ANYTHING). The accuracies for each PKU value are estimated by the scores given by its most related Armo-RM attribute (detailed in supplementary material (Holgado-Sánchez, Billhardt, and Ossowski 2026)).

atively small gap here is likely due to the regularization imposed on the grounding function implementation, given by the use of both ℓ_2 regularization and the centering parameter r .

Regarding grounding accuracy, VSL-RM and SEQ-RM achieve comparable results, consistent with the training curves (differences are not statistically significant⁷). Notably, Armo-RM outperforms all methods in ULTRAFEEDBACK, which may be attributed to its per-value training and the use of a MSE loss over quantitative annotations. In contrast, our methods are restricted to less informative qualitative labels, and the use of cross-entropy losses that are prone to overfitting. This limitation could be mitigated in future work, e.g. by label smoothing (Zhu, Jordan, and Jiao 2024).

Lastly, in Table 3, we analyze the value system weights predicted by VSL-RM and SEQ-RM⁸. Both models learned broadly similar value system representations, assigning comparable importance to the values. This indicates that the simultaneous learning procedure in VSL-RM does not substantially distort the learned value system, compared to a grounding implemented by a reward vector learned without a bias to maximize VSA. Notably, however, VSL-RM yields less “radical” weight distributions than SEQ-RM: it maintains a minimum weight of approximately 0.082 across all values, whereas SEQ-RM assigns some weights below 0.05. We consider that learning a value system model for a

whole dataset that does not neglect any value achieves a better alignment with the hypotheses of the dataset creators, as they would expect that all values have a meaningful contribution towards explaining the overall preferences based on their proposed values. Our learned grounding functions and value system weights with VSL-RM, are, in combination, qualitatively better in this regard.

As a conclusion, we have shown the advantages of an approach that defends an “intentional” interpretation of value alignment, by learning not only accurate approximations of value groundings but also ones which facilitate their combination through value systems as a means for an accurate estimation of value-based preferences. This is a fundamental advantage of our approach that is neglected by most existing methods that learn reward components and user preferences separately or sequentially, an approach that often leads to reward hacking in the fine-tuning phase (Zhou et al. 2024).

6 Conclusions and Future Work

In this paper, we have put forward a principled approach to value system learning for generative AI applications. Specifically, we have introduced a learning algorithm that, based on preference data, simultaneously learns representations of the value alignment (groundings) of prompt-response pairs (training a multi-objective reward model), and of their alignment with a certain value system (approximated by value importance weights). We have tested our solution against a single-objective reward learning baseline, a sequential variant of our algorithm, and a state-of-the-art related algorithm. We have shown that our algorithm prioritizes learning accurate value grounding models to serve as a correct basis

⁷In supplementary material (Holgado-Sánchez, Billhardt, and Ossowski 2026), we performed one-sided Welch t-tests to compare the metrics achieved by both methods.

⁸We omit the weights learned by BT-RM, as they are not based on valid value interpretations.

Method	HF	HO	TF	IF	PF	OB	CL	IR	SF
VSL-RM	0.096 ±0.002	0.347 ±0.008	0.408 ±0.007	0.147 ±0.003	0.223 ±0.003	0.125 ±0.005	0.368 ±0.005	0.203 ±0.003	0.082 ±0.002
SEQ-RM	0.050 ±0.009	0.477 ±0.010	0.316 ±0.006	0.158 ±0.006	0.265 ±0.003	0.110 ±0.006	0.410 ±0.004	0.168 ±0.001	0.047 ±0.001

Table 3: Average and standard deviation of value system weights of VSL-RM and SEQ-RM in ULTRAFeEDBACK (left) and PKU-ALIGN-ANYTHING (right). HF = Helpfulness, HO = Honesty, TF = Truthfulness, IF = Instruction Following, PF = Prompt Following, OB = Objectivity, CL = Clarity, IR = Information Richness, SF = Safety.

for value system estimation, and thus fosters interpretability, with minor losses in capturing overall value-based preferences (value systems).

Despite these positive results, our work also presents some limitations. First, even though our approach allows for explaining outcomes in terms of grounding functions and value system weights, interpretability is still limited by the fact that our algorithm performs an implicit aggregation of preferences regarding values and value systems of possibly divergent opinions, potentially obscuring minorities. Moreover, our current value system representation uses static linear weights, which in certain settings may put limits to the accuracy of modelling real value-based preferences.

In future work, we will address these potential shortcomings without sacrificing interpretability. For this purpose, we will introduce context-dependency into the value system model. A promising option is to dynamically switch between different linear value systems based on contextual conditions. Another major line of work will address methods to learn diverse value systems for heterogeneous societies in datasets with known annotation sources. Alternative datasets that we will explore for either task are OpenAssistant (Köpf et al. 2023) and PRISM (Kirk et al. 2024). Still, the former involves tree-like conversations, and thus requires an extension of our present model which is geared towards single prompt-response preferences. The latter appears particularly suitable for studying the alignment of our approach with sociological findings.

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. Special thanks to the faculty ETSII at Rey Juan Carlos University for providing the computational resources used in the experiments.

References

Adams, J.; Hu, B.; Veenhuis, E.; Joy, D.; Ravichandran, B.; Bray, A.; Hoogs, A.; and Basharat, A. 2025. Steerable Plu-

ralism: Pluralistic Alignment via Few-Shot Comparative Regression. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 15–25.

Anderson, M.; Anderson, S. L.; and Armen, C. 2006. An approach to computing ethics. *IEEE Intelligent Systems*, 21: 56–63.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Shawk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022a. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022b. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bose, A.; Xiong, Z.; Chi, Y.; Du, S. S.; Xiao, L.; and Fazel, M. 2025. LoRe: Personalizing LLMs via Low-Rank Reward Modeling. In *Second Conference on Language Modeling*.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*, 3(1): 1–122.

Boyd, S.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4): 324–345.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Manocha, D.; Huang, F.; Bedi, A.; and Wang, M. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Proc. 41st Int. Conf. on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 6116–6135. <https://www.jmlr.org>.

- Chen, D.; Chen, Y.; Rege, A.; Wang, Z.; and Vinayak, R. K. 2025. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment. In *The Thirteenth International Conference on Learning Representations*.
- Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4302–4310. Red Hook, NY, USA: Curran Associates Inc.
- Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377.
- Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2024. SAFE RLHF: SAFE REINFORCEMENT LEARNING FROM HUMAN FEEDBACK. In *12th International Conference on Learning Representations, ICLR 2024*. Red Hook, NY, USA: Curran Associates Inc. 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.
- Dognin, P.; Rios, J.; Luss, R.; Sattigeri, P.; Liu, M.; Padhi, I.; Riemer, M.; Nagireddy, M.; Varshney, K.; and Bouneffouf, D. 2025. Contextual Value Alignment. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.
- Dong, Y.; Wang, Z.; Sreedhar, M.; Wu, X.; and Kuchaiev, O. 2023. SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11275–11288. Singapore: Association for Computational Linguistics.
- Eisenstein, J.; Nagpal, C.; Agarwal, A.; Beirami, A.; D’Amour, A.; Dvijotham, D.; Fisch, A.; Heller, K.; Pfohl, S.; Ramachandran, D.; Shaw, P.; and Berant, J. 2024. Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate Reward Hacking. arXiv:2312.09244.
- Fallah, A.; Ozdaglar, A.; and Pattathil, S. 2020. An Optimal Multistage Stochastic Gradient Method for Minimax Problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 3573–3579.
- Guo, Y.; Cui, G.; Yuan, L.; Ding, N.; Sun, Z.; Sun, B.; Chen, H.; Xie, R.; Zhou, J.; Lin, Y.; Liu, Z.; and Sun, M. 2024. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1437–1454. Miami, Florida, USA: Association for Computational Linguistics.
- Gupta, R.; Sullivan, R.; Li, Y.; Phatale, S.; and Rastogi, A. 2025. Robust Multi-Objective Preference Alignment with Online DPO. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26): 27321–27329.
- Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S. J.; and Dragan, A. 2017. Inverse Reward Design. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Holgado-Sánchez, A.; Billhardt, H.; and Ossowski, S. 2026. A Method for Learning Value Systems in Generative AI. (Full version with supplementary material), arXiv:2607.16903.
- Holgado-Sánchez, A.; Billhardt, H.; Ossowski, S.; and Degli-Esposti, S. 2025. Learning the Value Systems of Societies from Preferences. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, 1123–1130. IOS Press.
- Holgado-Sánchez, A.; Billhardt, H.; Fernández, A.; and Ossowski, S. 2026. Learning the value systems of agents with preference-based and inverse reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 40(4): 4.
- Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2024. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*. <https://openreview.net>: OpenReview.
- Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023. BEAVERTAILS: towards improved safety alignment of llm via a human-preference dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*. Red Hook, NY, USA: Curran Associates Inc.
- Ji, J.; Zhou, J.; Lou, H.; Chen, B.; Hong, D.; Wang, X.; Chen, W.; Wang, K.; Pan, R.; Li, J.; Wang, M.; Dai, J.; Qiu, T.; Xu, H.; Li, D.; Chen, W.; Song, J.; Zheng, B.; and Yang, Y. 2024. Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback. arXiv:2412.15838.
- Karanik, M.; Billhardt, H.; Fernández, A.; and Ossowski, S. 2024. On the relevance of value system structure for automated value-aligned decision-making. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 679–686. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702433.
- Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024. The PRISM alignment dataset: what participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.
- Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z.-R.; Stevens, K.; Barhoum, A.; Duc, N. M.; Stanley, O.; Nagyfi, R.; ES, S.; Suri, S.; Glushkov, D.; Dantuluri, A.

- Maguire, A.; Schuhmann, C.; Nguyen, H.; and Mattick, A. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.
- Lera-Leri, R. X.; Liscio, E.; Bistaffa, F.; Jonker, C. M.; López-Sánchez, M.; Murukannaiah, P. K.; Rodríguez-Aguilar, J. A.; and Salas-Molina, F. 2024. Aggregating value systems for decision support. *Knowledge-Based Systems*, 287: 111453.
- Li, C.; Zhang, H.; Xu, Y.; Xue, H.; Ao, X.; and He, Q. 2025. Gradient-Adaptive Policy Optimization: Towards Multi-Objective Alignment of Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11214–11232. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.
- Liscio, E.; Lera-Leri, R.; Bistaffa, F.; Dobbe, R. I. J.; Jonker, C. M.; López-Sánchez, M.; Rodríguez-Aguilar, J. A.; and Murukannaiah, P. K. 2023. Value Inference in Sociotechnical Systems. In *Proc. AAMAS, Blue Sky Ideas Track*, 1774 – 1780.
- Liscio, E.; van der Meer, M.; Siebert, L. C.; Jonker, C. M.; Mouter, N.; and Murukannaiah, P. K. 2021. Axes: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, 799–808. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.
- Montes, N.; Osman, N.; Sierra, C.; and Slavkovik, M. 2023. Value Engineering for Autonomous Agents. arxiv:2302.08759.
- Montes, N.; and Sierra, C. 2022. Synthesis and properties of optimally value-aligned normative systems. *Journal of Artificial Intelligence Research*, 74: 1739–1774.
- Osman, N.; and d'Inverno, M. 2024. A Computational Framework of Human Values. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '24, 1531–1539. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.
- Ramé, A.; Couairon, G.; Shukor, M.; Dancette, C.; Gaya, J.-B.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.
- Reddy, S.; Dragan, A.; Levine, S.; Legg, S.; and Leike, J. 2020. Learning Human Objectives by Evaluating Hypothetical Behavior. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8020–8029. <https://www.jmlr.org>: JMLR.org.
- Rodríguez-Soto, M.; Rădulescu, R.; Bistaffa, F.; Ricart, O.; Mayoral-Macau, A.; López-Sánchez, M.; Rodríguez-Aguilar, J. A.; and Nowé, A. 2026. Multi-objective reinforcement learning for provably incentivising alignment with value systems. *Artificial Intelligence*, 351: 104460.
- Rodríguez-Soto, M.; Serramia, M.; López-Sánchez, M.; and Rodríguez-Aguilar, J. A. 2022. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24: 9.
- Russell, S. 2022. Artificial Intelligence and the Problem of Control. In Werthner, H.; Prem, E.; Lee, E. A.; and Ghezzi, C., eds., *Perspectives on Digital Humanism*, 19–24. Springer.
- Schwartz, S. H. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, 1–65. Elsevier.
- Serramia, M.; López-Sánchez, M.; Rodríguez-Aguilar, J. A.; Rodríguez, M.; Wooldridge, M.; Morales, J.; and Ansótegui, C. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, 1294–1302. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
- Siebert, L. C.; Liscio, E.; Murukannaiah, P. K.; Kaptein, L.; Spruit, S.; Hoven, J. V. D.; and Jonker, C. 2022. Estimating Value Preferences in a Hybrid Participatory System. *Frontiers in Artificial Intelligence and Applications*, 354: 114 – 127.
- Sinnott-Armstrong, W. 2023. Consequentialism. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.
- Soares, N. 2018. *The Value Learning Problem*, 89–97. Chapman and Hall/CRC.
- Sorensen, T.; Jiang, L.; Hwang, J. D.; Levine, S.; Pyatkin, V.; West, P.; Dziri, N.; Lu, X.; Rao, K.; Bhagavatula, C.; Sap, M.; Tasioulas, J.; and Choi, Y. 2024. Value kaleidoscope: engaging AI with pluralistic human values, rights, and duties. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN 978-1-57735-887-9.
- Sumers, T.; Hawkins, R.; Ho, M. K.; Griffiths, T.; and Hadfield-Menell, D. 2022. How to talk so AI will learn: Instructions, descriptions, and autonomy. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 35 of *NIPS '22*, 34762–34775. Red Hook, NY, USA: Curran Associates, Inc.
- Tennant, E.; Hailes, S.; and Musolesi, M. 2025. Moral Alignment for LLM Agents. In *The Thirteenth International Conference on Learning Representations*.

- Vamplew, P.; Dazeley, R.; Foale, C.; Firmin, S.; and Mumery, J. 2018. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1): 27–40.
- Van Moffaert, K.; Drugan, M.; and Nowé, A. 2013. Scalarized multi-objective reinforcement learning: novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 191–199. IEEE.
- Wang, H.; Lin, Y.; Xiong, W.; Yang, R.; Diao, S.; Qiu, S.; Zhao, H.; and Zhang, T. 2024a. Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8642–8655. Bangkok, Thailand: Association for Computational Linguistics.
- Wang, H.; Xiong, W.; Xie, T.; Zhao, H.; and Zhang, T. 2024b. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10582–10592. Miami, Florida, USA: Association for Computational Linguistics.
- Wang, X.; Le, Q.; Ahmed, A.; a Enmao Diao; Zhou, Y.; Baracaldo, N.; Ding, J.; and Anwar, A. 2025. MAP: Multi-Human-Value Alignment Palette. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 81284–81313. Red Hook, NY, USA: Curran Associates Inc.
- Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J. J.; Sreedhar, M. N.; and Kuchaiev, O. 2024c. HelpSteer2: open-source dataset for training top-performing reward models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.
- Yang, R.; Pan, X.; Luo, F.; Qiu, S.; Zhong, H.; Yu, D.; and Chen, J. 2024. Rewards-in-context: multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. <https://www.jmlr.org>: JMLR.org.
- Zamani, M.; Abbaszadehpeivasti, H.; and de Klerk, E. 2022. Convergence rate analysis of the gradient descent-ascent method for convex-concave saddle-point problems. arXiv:2209.01272.
- Zhou, Z.; Liu, J.; Shao, J.; Yue, X.; Yang, C.; Ouyang, W.; and Qiao, Y. 2024. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 10586–10613. Bangkok, Thailand: Association for Computational Linguistics.
- Zhu, B.; Jordan, M. I.; and Jiao, J. 2024. Iterative data smoothing: mitigating reward overfitting and overoptimization in RLHF. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Ziegfeld, L.; Kox, E.; Akrum, I.; and Heijnen, M. 2025. Aligning AI Systems with Human Values: A Method for Identifying and Specifying Values. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2821–2831.