Journal of Artificial Intelligence Research 82 (2025) 819-850 Submitted 06/2023; published 02/2025

Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems

Enrico Liscio E.LISCIO@TUDELFT.NL
Luciano C. Siebert L.CAVALCANTESIEBERT@TUDELFT.NL
Delft University of Technology, the Netherlands

Catholijn M. Jonker C.M.JONKER@TUDELFT.NL
Delft University of Technology, the Netherlands and *Leiden University, The Netherlands*

Pradeep K. Murukannaiah P.K.MURUKANNAIAH@TUDELFT.NL
Delft University of Technology, the Netherlands

Abstract

Understanding citizens' values in participatory systems is crucial for citizen-centric policy-making. We envision a hybrid participatory system where participants make choices and provide motivations for those choices, and AI agents estimate their value preferences by interacting with them. We focus on situations where a conflict is detected between participants' choices and motivations, and propose methods for estimating value preferences while addressing detected inconsistencies by interacting with the participants. We operationalize the philosophical stance that "valuing is deliberatively consequential." That is, if a participant's choice is based on a deliberation of value preferences, the value preferences can be observed in the motivation the participant provides for the choice. Thus, we propose and compare value preferences estimation methods that prioritize the values estimated from motivations over the values estimated from choices alone. Then, we introduce a disambiguation strategy that combines Natural Language Processing and Active Learning to address the detected inconsistencies between choices and motivations. We evaluate the proposed methods on a dataset of a large-scale survey on energy transition. The results show that explicitly addressing inconsistencies between choices and motivations improves the estimation of an individual's value preferences. The disambiguation strategy does not show substantial improvements when compared to similar baselines—however, we discuss how the novelty of the approach can open new research avenues and propose improvements to address the current limitations.

1. Introduction

Values, spanning concepts such as self-determination and sustainability, are the standards or criteria that justify one's opinions and actions and are intrinsically linked to goals (Schwartz, 2012). Values form an ordered system of priorities and the relative importance one ascribes to values (one's *value preferences*) guides action. Since values define shared goals and are essential for human cooperation, they are deemed critical to developing AI that can integrate beneficially into our society (Russell et al., 2015; Gabriel, 2020). Yet, how individuals ascribe relative priorities among values can vary significantly across people, socio-cultural environments (Dignum, 2017), and decision contexts (Hill & Lapsley, 2009). Identifying and reasoning about individuals' value preferences has been recognized as the challenge of *value inference* (Liscio et al., 2023), encompassing AI and hybrid human-AI methods proposed

to identify the values relevant to a decision-making process (Wilson et al., 2018; Liscio et al., 2022) or to detect values in language (Kiesel et al., 2022; Qiu et al., 2022). Such semi-automated approaches offer the chance to infer individuals' values at a large scale.

One crucial field that can benefit from large-scale value inference is policy-making. Enhancing citizen participation in decision-making processes is high on the European policy agenda (Dallhammer et al., 2018). Initiatives to foster citizens' political power and engagement have been proposed through the use of digital platforms for participatory decision-making (Lafont, 2015; Mouter et al., 2021) and deliberation (Friess & Eilders, 2015; Iandoli et al., 2016; Shortall et al., 2022). To this end, eliciting stakeholders' preferences over competing alternatives only provides superficial information on the debate. Instead, considering stakeholders' values on a decision-making subject is crucial for crafting long-term policies on the subject (Miller, 2016) since values preferences tend to be stable over time (Schwartz, 2012). For instance, consider a policy-maker drafting subsidy strategies for solar panels; knowing what value trade-offs motivated the citizens (e.g., sustainability vs. economic efficiency) will inform long-term solutions as well as similar decisions in the future.

Within the value inference process, *value preferences estimation* refers to the challenge of estimating an individual's preferences over a given set of relevant values[1] (Liscio et al., 2023). Estimating value preferences on an individual level (as opposed to a population level) allows for (1) a detailed understanding of how different individuals prioritize values; (2) interactive approaches for disambiguation on an individual level. To inform the policy-maker on the population's preferences, value preferences can then be later aggregated at the collective level (e.g., Lera-Leri et al. (2024)).

Value preferences estimation has been traditionally performed based on one's *choices* over competing alternatives, e.g., from answers to value surveys (Schwartz, 2012; Graham et al., 2013) or from one's action in a context (Liscio et al., 2023). In other words, estimating value preferences involves identifying or defining, e.g., through expert assessment or bottom-up aggregation, the relationship between an individual's choice and a value, such as whether the choice promotes the value. However, estimating one's value preferences can be challenging due to the intrinsic uncertainty in defining value-choices relationships and the ambiguity that multiple value preferences could possibly explain a choice (Mindermann & Armstrong, 2018). Estimating value preferences from both one's choices in a context and the verbal *motivations* for supporting these choices provides additional insights that could not be achieved considering only one source of information. For instance, consider an individual who recently installed solar panels; they may have been motivated by the values of sustainability or economic efficiency, or both, or neither. Seeking verbal motivations for their choices might unveil their value preferences.

We envision a semi-automated approach to value preferences estimation, where AI agents, supported by natural language processing (NLP) techniques, interpret the motivations provided by the participants in support of their choices, and combine the information contained in choices and motivations to estimate their value preferences. But what if the

1. Value preferences estimation falls within the realm of descriptive ethics (Hämäläinen, 2016), aimed at discerning the guiding principles of individuals (with the assumption that they will aim to choose actions that align with their preferred values). It is important to distinguish this from normative ethical theories, such as deontology or utilitarianism, which prescribe how rational agents *ought to* behave. These theories focus on moral decision-making principles and mechanisms, rather than individual value preferences.

information extracted from the choices conflicts with the information extracted from the motivations given in support of those choices? To target such conflicts, we propose a hybrid intelligence (HI) (Akata et al., 2020) approach where value preferences are estimated through the combination of artificial and human intelligence.

Consider the aforementioned scenario where citizens provide their choices for subsidy strategies for solar panels. Values such as sustainability and economic efficiency are relevant factors that might influence individuals' decisions to install solar panels in their homes. Let us assume that an individual supports subsidy policy A, for which the value of sustainability was deemed relevant (e.g., by looking into previous decisions and motivations from other individuals, or through expert input). This choice alone does not necessarily reveal their underlying reasons. However, if they solely mention economic efficiency when motivating policy A, but not sustainability, a conflict arises. We target conflicts between choices and motivations through value preferences *estimation* and *disambiguation*, as shown in Figure 1.

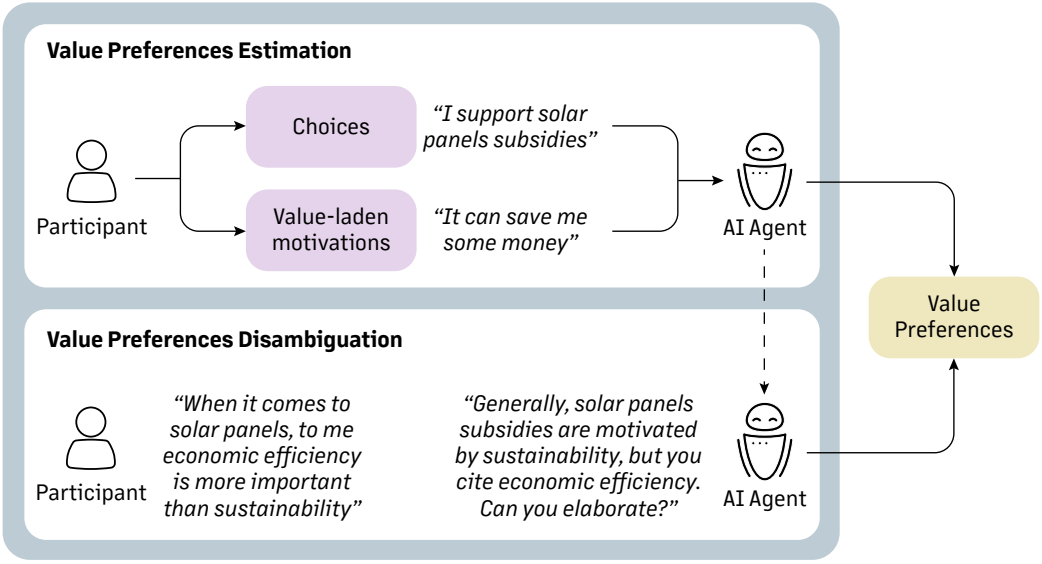


Figure 1: A hybrid participatory system where human participants make choices and motivate those choices, and AI agents estimate and disambiguate participants' value preferences.

We propose and compare five methods for estimating value preferences from the choices and motivations provided by participants in a participatory system. These methods combine participants' choices and motivations in different ways. First, we explore methods that only consider the choices or the motivations. Next, we propose three methods that employ a combination thereof. When choice-motivation conflicts arise, these methods follow the philosophical account that "valuing is deliberatively consequential"[2] (Scheffler, 2012), i.e., if one's choice is based on a deliberation of value preferences, the value preferences can be observed in the motivation provided for the choice (Dietz & Stern, 1995; Kenter et al., 2016;

2. In this context, 'consequential' refers to one's disposition to treat certain value-related considerations as a reason for action. It should not be confused with 'consequentialism' as a normative ethical theory that asserts the consequences of one's actions should be the primary basis for moral judgments.

Pigmans et al., 2019). Thus, the methods prioritize the values observed in the motivations over those observed in the choices.

Nevertheless, the detected choice-motivation conflicts ought to be addressed. Such conflicts may be caused by (1) mistakes in the value preferences estimation process (e.g., misclassification of the values supporting the participants' motivations by an NLP model), or (2) genuine inconsistencies between the participants' choices and motivations, e.g., due to participants having different assumptions regarding values that drive a choice, or due to the value-action gap (Franco & Ghisetti, 2022). In both cases, addressing the inconsistencies can be beneficial. If the inconsistency is caused by a mistake in the automatic value preferences estimation process, the involved participant should be asked to resolve the mistake, e.g., by correcting a misclassification of the NLP model. In case the interpretation is confirmed to be correct and the inconsistency is accurately detected, the participant can be guided through a process of self-reflection (Lim et al., 2019; Liscio et al., 2023) and offered the chance to change their choices or provide additional motivations.

In participatory systems, not all participants may be available to take part in such interactions, and the required additional effort may dissuade participants from engaging (Shortall et al., 2022). Inspired by Active Learning (AL) (Settles, 2012), we propose a disambiguation strategy that guides the interactions between AI agents and participants, following the rationale that, by addressing the most informative participants first, the quality of value preferences estimation should rapidly improve for all participants. Accordingly, the strategy iteratively selects the participants whose value preferences estimated solely from their choices are most different from the value preferences estimated solely from their motivations. We test this strategy by retrieving the correct interpretation of the motivations provided by the selected participants (i.e., the correct values that support their motivations) to iteratively improve the NLP model tasked to predict the values that support the participants' motivations, which are in turn used to estimate their value preferences.

Contribution We propose a method for estimating individuals' value preferences through a disambiguation strategy. Our method is composed of two parts. First, we propose and compare five methods for estimating individual value preferences. We employ the proposed methods to estimate the value preferences of the participants in a large-scale survey on energy transition (Itten & Mouter, 2022), where participants allocate a number of points among policy options and textually motivate their choices. We evaluate the extent to which our methods' estimations concur with those of human evaluators. Our results show that addressing the inconsistencies between choices and motivations improves the estimation of value preferences. Second, we propose and evaluate a disambiguation strategy that is driven by inconsistencies between participants' choices and motivations. We evaluate the strategy in an active learning setting with the value-annotated survey on energy transition, and compare it to traditional NLP AL strategies. We show that our method leads to comparable results to the tested baselines, both in NLP performance and value preferences estimation. We discuss the results and elaborate on future directions.

Extension This paper extends the conference paper from Siebert et al. (2022), where we propose and compare five methods for estimating value preferences from one's choices and motivations. We extend the work by introducing the disambiguation strategy, which naturally complements the value preferences estimation methods in two ways. First, it relies

on the same philosophical account by addressing inconsistencies between one's choices and motivations. Second, it situates the value preferences estimation endeavor in a realistic setting, by proposing a concrete and scalable approach for performing value preferences estimation *during* a participatory process. We extend the evaluation in the original paper to validate our proposed disambiguation strategy by using it as a sampling strategy in an AL setting and comparing it to traditional sampling strategies.

Organization Section 2 discusses related works. Section 3 introduces the context behind our analyses. Section 4 describes the methods we propose for value preferences estimation and for the disambiguation strategy. Section 5 describes our experimental setup and Section 6 presents our results. Finally, Section 7 discusses the limitations of our work and Section 8 concludes the paper.

2. Related Works

We discuss related works on valuing, estimating value preferences, NLP techniques for classifying values from text, and active learning.

2.1 Valuing

Smith et al. (1989) describe valuing as a form of desiring. However, this conception is limited. For instance, to value someone's leadership skills does not mean desiring this person's leadership skills for oneself, but recognizing them as something positive. Thus, some philosophers have rejected the reduction of valuing to desiring and proposed that it should be perceived in a broader sense, as having a favorable attitude towards something, involving both reason and emotion (Scheffler, 2012; Frankfurt, 2018).

Scheffler (2012) suggests that having such a favorable attitude towards something implies a deliberative significance. That is, to value X involves not only seeing X as a source of reasons for action (a view also supported by Schwartz (2012)) but also having considerations related to X in a relevant context. For example, if one values *privacy*, they would contemplate in relevant contexts to treat considerations about the impact of proposed actions on their privacy as having deliberative importance. In other words, Scheffler states that valuing is *deliberatively consequential*.

Several researchers support this view. For instance, Dietz and Stern (1995) consider the notion that people assess their options in terms of expected outcomes (subjective expected utility model), referring to personal values. Kenter et al. (2016) propose the Deliberative Value Formation model, in which deliberation is considered to form values through processes that may inform and enable reflection. In the context of citizen participation, Pigmans et al. (2019) suggest that, if values that stakeholders perceive as relevant can be identified as part of the deliberation process, reflection and mutual understanding could be promoted.

In this work, we follow Scheffler (2012) and others by taking the stance that, if one's choice is based on a deliberation of value preferences, the value preferences can be observed

in the motivation provided for the choice. This approach can support increasing legitimacy in decision-making, by providing a grounded approach for estimating value preferences.[3]

2.2 Estimating Value Preferences

Survey instruments such as the Portrait Value Questionnaire (Schwartz, 2012), Schwartz Value Survey (Schwartz, 2012), Value Living Questionnaire (Wilson et al., 2010), and Moral Foundations Questionnaire (Graham et al., 2011) have been used to estimate an individual's preferences towards a set of values. Further, some approaches combine self-reported surveys with participatory design (Pommeranz et al., 2012; Liao & Muller, 2019), following the principles of Value Sensitive Design (Friedman & Hendry, 2019). However, value questionnaires have been criticized for being incomplete and not context-sensitive (Le Dantec et al., 2009; Boyd et al., 2015). In this work, we do not query participants directly about their value preferences, but evaluate their choices and related motivations in context.

Alternatively, value preferences can be estimated from a bottom-up approach by analyzing human behavior and choices. In the field of economics, values have been elicited via revealed preference methods such as direct elicitation and multiple price lists (Benabou et al., 2020). For complex and high-dimensional environments, inverse reinforcement learning algorithms (Ng & Russell, 2000), which focus on extracting a "reward function" given observed optimal behavior, show promising results (Russell, 2019). However, critiques on the infeasibility of estimating an individual's rationality and preferences (including value preferences) simultaneously (Mindermann & Armstrong, 2018) suggest the need for additional normative assumptions, e.g., an explicit model of the cognitive processes that guided a given behavior. Furthermore, the use of reward or objective functions has been argued not to be well-suited for modeling human values or other normative concepts (i.e., judgments of what is right, wrong, good, or bad) (Eckersley, 2019). We seek to address such critiques by (1) incorporating textual motivations provided by humans for their choices and using NLP approaches to automatically classify the values that underlie the motivations and (2) using partially ordered preferences for modeling value preferences.

2.3 Classifying Values from Text

A classical approach to value classification from text is through value dictionaries—lists of word characteristic of certain values—by measuring the relative frequency of the words describing each value (Pennebaker et al., 2001) e.g., the Moral Foundation Dictionary (Graham et al., 2013). These dictionaries have been expanded through semi-automated methods (Wilson et al., 2018; Araque et al., 2020; Hopp et al., 2021) or through NLP techniques (Ponizovskiy et al., 2020; Araque et al., 2021), and limitations related to word count techniques have been approached via word embedding models (Garten et al., 2018; Bahgat et al., 2020; Pavan et al., 2020). More recent approaches use supervised machine learning (Liscio et al., 2022; Kiesel et al., 2022; Alshomary et al., 2022; Huang et al., 2022; Liscio et al., 2023; van der Meer et al., 2023; Park et al., 2024), where NLP models are trained on datasets annotated with value taxonomies, such as the Moral Foundation Twitter Corpus

3. In this work, we do not aim to model individual moral decision-making, e.g., as discussed by Haidt (2001), who argues that moral choices are based on intuitions rather than reasoning or deliberation. Instead, we focus on valuing as a deliberative process to support and legitimize participation.

(Hoover et al., 2020) and ValueNet (Qiu et al., 2022). Our method builds on this approach, as we train an NLP model on an annotated dataset. However, we expand on the literature by employing an active learning approach.

2.4 Active Learning

The key idea behind Active Learning (AL) is that a supervised ML algorithm can achieve good performance with few training examples if such examples are suitably selected (Settles, 2012). In a traditional AL setting, a large set of *unlabeled data* is available, and an *oracle* (e.g., human annotators) can be consulted to annotate the unlabeled data. A *sampling strategy* is used to iteratively select the next batch of unlabeled data to be annotated by the oracle, with the intent of rapidly improving the performance of the ML algorithm. A commonly used sampling strategy is uncertainty sampling (Ren et al., 2021), where at every iteration the ML algorithm is used to predict labels on all the unlabeled data, and the m unlabeled data with the highest label entropy are selected as the next batch to be annotated (i.e., the data on which the model is least confident about its prediction).

AL has been extensively used in NLP applications (Zhang et al., 2022), with two main strategy approaches. On the one hand, some strategies use the *informativeness* of each unlabeled instance individually, e.g., by measuring the uncertainty of the prediction or the norm of the gradient (Zhang et al., 2017). The unlabeled instances that are estimated to be most informative are selected to be labeled by the oracle. On the other hand, other strategies focus on the *representativeness* of the data, e.g., by selecting data points that are most representative of the unlabeled set (Zhao et al., 2020) or that are most different from the data that is already labeled (Erdmann et al., 2019). In general, state-of-the-art AL strategies exploit information about the NLP task (i.e., about the NLP model and the available data) with the intent of rapidly improving the performance of the NLP model. However, in our setting, the NLP model is a means to the end of estimating value preferences. Hence, we propose a strategy that is driven by the informativeness of the unlabeled data, but where the informativeness is derived by the downstream task of value preferences estimation.

3. Background

We introduce the dataset and formalize the key concepts to provide a background for our methods and experiments.

3.1 Participatory Value Evaluation (PVE)

We estimate individual value preferences from choices and motivations provided via Participatory Value Evaluation (PVE) (Mouter et al., 2021), an online participatory system. We use data from a PVE conducted between April and May 2020 involving 1376 participants (Itten & Mouter, 2022), aimed at supporting the municipality of Súdwest-Fryslân in the Netherlands in co-creating an energy transition policy, increasing citizen participation, and avoiding public resistance as happened in previous projects on sustainable energy (Haag, 2019). The main question to the citizens was: "What do you find important for future decisions on energy policy?" Six policy options (Table 1) were developed in consultation with

45 citizens. These options were presented in the PVE platform, with the participants asked to distribute 100 points among them. In most cases, participants assigned points to more than one option, with options o_1 and o_2 receiving more than half of the points on average. After dividing the points, the participants had the chance to provide a textual motivation in support of each of the options to which they had allocated points. 876 participants provided at least one motivation for their choices, resulting in a total of 3229 motivations.

Policy option	Description	Avg. points distributed
o_1	The municipality takes the lead and unburdens you	29.05
o_2	Inhabitants do it themselves	21.72
o_3	The market determines what is coming	9.39
o_4	Large-scale energy generation will occur in a small number of places	15.01
o_5	Betting on storage (Súdwest-Fryslân becomes the battery of the Netherlands)	12.96
o_6	Become a major energy supplier in the Netherlands	4.71

Table 1: Policy options available in the energy transition PVE.

The motivations were annotated with the underlying values as part of the original data collection. We refer to Kaptein (2020) for a detailed description of the annotation procedure, which we summarize here. The values embedded in the textual motivations were identified by a team of four annotators using a grounded theory approach (Heath & Cowley, 2004). The annotators were first introduced to foundational concepts (Schwartz, 2012; Graham et al., 2013) and examples of values. Then, they were asked to annotate any keywords from the motivations that relate to values. After a consolidation round, annotators agreed on a list with 18 values (as presented in Appendix A.1). In this paper, we consider only the most frequent values (values mentioned at least 250 times across all project options) to demonstrate our methods. This allows us to perform an in-depth analysis and provide a digestible overview of the results while managing the computational load. Nevertheless, our methods are agnostic of the number of values, as we further discuss in Section 7. Table 2 shows the value list we consider in our experiments.

Value ID	Value name	Description
v_1	Cost-effectiveness	Money must be well spent and the project must be profitable. No waste. Costs should not be too high
v_2	Nature and landscape	Nature and environment are important. Horizon pollution is often seen as negative. Preserving the Frisian landscape is central
v_3	Leadership	Clarity and control over the sustainability of the energy system. Often about an organization or person that has to take charge
v_4	Cooperation	Working together on a goal. Residents can work together, but also groups and organizations
v_5	Self-determination	The opportunity for residents to make their own decision on renewable energy and to be able to implement it

Table 2: Considered values for the energy transition PVE.

Table 3 shows the number of annotations provided for each of the values we analyze (described in Table 2). Although all values have more than 250 annotations (our selection criterion), these values were not annotated equally across the choice options. For example, v_3 was annotated 349 (∼76%) times for o_3, and only 3 times for o_6.

Annotated values	Options						
	o_1	o_2	o_3	o_4	o_5	o_6	O
v_1	90	85	102	85	89	58	**509**
v_2	50	29	11	269	27	47	**433**
v_3	349	40	42	13	11	3	**458**
v_4	80	131	35	17	13	31	**307**
v_5	35	305	7	8	20	16	**391**
V	**604**	**590**	**197**	**392**	**160**	**155**	

Table 3: Distribution of values annotated for each policy option.

3.2 Formalization

We formalize the concepts associated with the PVE (choices and motivations) and with value preferences estimation (value systems and value-option matrix). These concepts are related as shown in Figure 2.

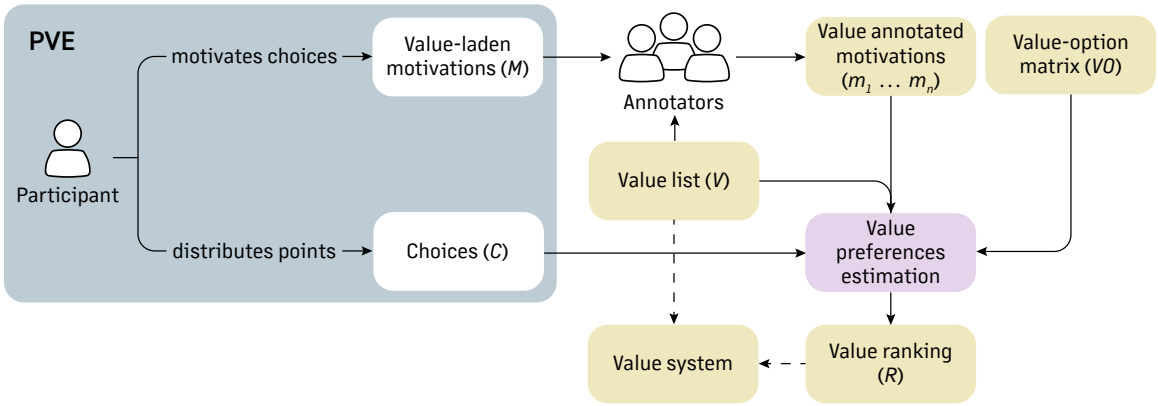


Figure 2: Each PVE participant makes choices C (i.e., distributes points to the policy options) and provides motivations M to their choices. The participant's value system is defined as the ranking R over a set of values V. Our proposed value preferences estimation methods estimate R based on (1) a given value list V, (2) the choices C, (3) the values annotated in the motivations, and (4) an initial estimate of their value-option matrix VO.

3.2.1 Value System

Values can be ordered according to their subjective importance as guiding principles (Schwartz, 2012). Each person has a *value system* that internally defines the importance the values have to a person according to their preference and context. We represent this value prefer-

ence via a ranking (Zintgraf et al., 2018). Adapting from Serramia et al. (2021), we formally define a value system as follows.

Def 1 *A value system is a pair $\langle V, R\rangle$, where V is a non-empty set of values, and R is the ranking of V which represents a person's value preference.*

Def 2 *A ranking R of V is a reflexive, transitive, and total binary relation, noted as $v_a \succeq v_b$. Given $v_a, v_b \in V$, if $v_a \succeq v_b$, we say v_a is more preferred than v_b. If $v_a \succeq v_b$ and $v_b \succeq v_a$, then we note it as $v_a \sim v_b$ and consider v_a and v_b indifferently preferred. However, if $v_a \succeq v_b$ but it is not true that $v_b \succeq v_a$ (i.e., $v_a \neq v_b$), then we note it as $v_a \succ v_b$.*

In this work, we fix the set of values V for all participants (see Table 2) and we propose methods to estimate individuals' rankings over V. We refer to this task as *value preferences estimation* in the remainder of the paper. Further, ranking as defined here allows us to know the preferences between any pair of elements (unlike partial orders). We recognize that one's value preferences might not be a total order, since one could consider a given set of values incomparable. Yet, we focus on total orders as an initial step in estimating value preferences, given the challenges of fairly aggregating partial orders (Pini et al., 2005).

3.2.2 Choices and Motivations

Our goal is to estimate an individual i's value preferences via a ranking, R^i, from i's choices and the motivations provided for these choices. Let $O = \{o_1, \dots o_j, \dots o_n\}$ be a set of n options that i can choose from in a specific context (for example, the policy options presented in Table 1). We assume that i indicates their preferences, C^i, among the choices in O by distributing a certain number of points, p, among the options in O.

$$C^i = \{c_1, \dots c_j, \dots c_n\}, \quad c_j \in [0, p], \quad \sum c_j = p$$

Let M^i be the set of motivations that i provides for their choices:

$$M^i = \{m_1, \dots m_j, \dots m_n\}$$

Following the premise that valuing is deliberatively consequential, if an individual's value system influences their choice c_j, we expect them to mention the values that support choice c_j in the motivation provided. Thus, we represent a motivation m_j as the set of values (for example, a subset of the values in Table 2) that are mentioned in the motivation (with the set being empty if i assigned no points to o_j (i.e., $c_j = 0$) and thus no motivation was provided for that policy option):

$$m_j = \{v_1, \dots v_l, \dots v_m\}, \text{ if } v_l \in V \text{ influenced } c_j, \text{ with } m_j = \varnothing \text{ if } c_j = 0$$

3.2.3 Value-Option Matrix

We define a value-option matrix as follows:

Def 3 *An individual's i value-option matrix VO^i is a binary matrix with $|V|$ (number of values) rows and $|O|$ (number of options) columns, where:*

$$VO^i(v, o) = \begin{cases} 1, & \textit{if value } v \textit{ is relevant for option } o \\ 0, & \textit{otherwise.} \end{cases}$$

We employ VO^i as a fine-grained representation of an individual i's value preferences that displays which values are relevant for which option for that individual. In the following section, we describe how the proposed value preferences estimation methods employ and adjust VO^i to compute the individual's value ranking R^i.

4. Method

We propose a method for estimating value preferences through a disambiguation strategy. In this section, we present the two components of our method: the estimation of value preferences and the disambiguation strategy.

4.1 Value Preferences Estimation

Our goal is to estimate an individual's i value ranking R^i from the division of points across a set of *choices* and the textual *motivations* provided to each choice.

First, we propose two methods that compute R^i based either on i's motivations (method M, resulting in R^i_M) or on i's choices (method C, resulting in R^i_C). We employ these two methods as baselines. Next, we propose three methods that combine choices and motivations. Method TB (resulting in R^i_{TB}) resolves ties in R^i_C by using the motivations provided by i. Method MC (resulting in R^i_{MC}) and method MO (resulting in R^i_{MO}) update VO^i by addressing the inconsistencies between choices and motivations and between motivations provided for different policy options, respectively. Figure 3 shows the main elements of the five methods, which are described in detail in the remainder of this subsection. These methods can be applied sequentially—however, the order in which they are applied can change the final ranking. Furthermore, methods C, TB, MC, and MO take as input an initial estimate of VO^i (which gets updated in the case of MC and MO). We elaborate on the choice of the initial VO^i in our experiments in Section 5.1.

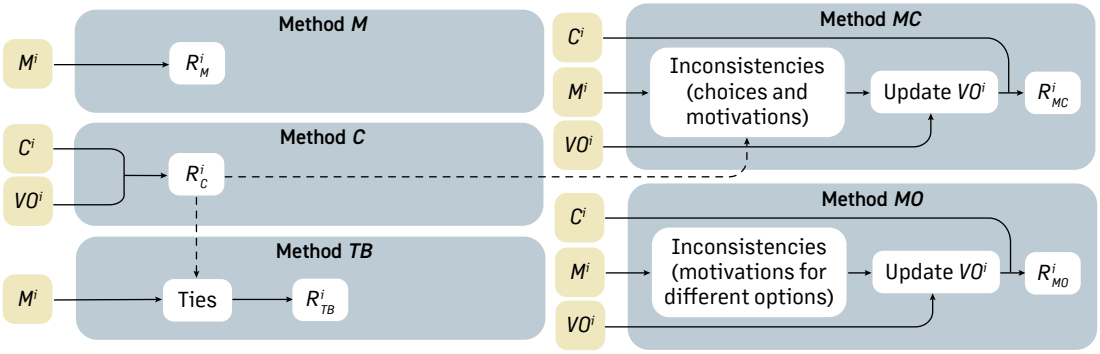


Figure 3: Overview of the five proposed value preferences estimation methods.

4.1.1 Method M

To estimate an individual's value ranking R^i_M solely based on the motivations M^i provided to their choices C^i, we first count how many times a given value is mentioned (i.e.,

annotated) in any of the motivations provided, and attribute one point to each time it is mentioned. Then, we infer the ranking R^i_M by ordering the values accordingly.

4.1.2 Method C

To estimate an individual's value ranking R^i_C solely based on their choices C^i (vector of size $|O|$, i.e., the number of options), we assume that the individual's choices completely align with their value preferences. First, we compute the importance of the values (U^i) for the individual by weighing the values supported by each option (o_j) with the points (c_j) the individual assigns to the option. Then, we infer a ranking R^i_C from U^i, by ordering the values in V according to their importance score in U^i.

$$U^i = VO^i \times {C^i}^T \tag{1}$$

$$R^i_C = \texttt{rank}(U^i) \tag{2}$$

where U^i is a V-sized vector of non-negative integers. For instance, assume that i distributes their points to the six policy options as $C^i = \{10, 20, 30, 20, 0, 20\}$ and the initial estimate of VO^i is as shown in Table 4. The multiplication of $VO^i \times {C^i}^T$ results in $U^i = \{100, 70, 60, 80, 30\}$. R^i_C is determined by ordering the values in V according to their importance score: $v_1 \succ v_4 \succ v_2 \succ v_3 \succ v_5$.

4.1.3 Method TB: Motivations as Tie Breakers

We use the motivations M^i as *tie breakers* to reduce indifferent preferences in a value ranking. We start with a given ranking R^i (e.g., R^i_C). Then, let us define that a tie $\tau_{a,b} \in R^i$ between two values $v_a, v_b \in V$ is present when v_a and v_b are indifferently preferred ($v_a \sim v_b$). If there is a tie $\tau_{a,b}$ and if at least one of the motivations mentions v_a but none of the motivations mention v_b, then the TB method considers $v_a \succ v_b$, and thus breaks the tie. If both values are mentioned in one of the motivations or not mentioned in any motivation, the tie remains. Algorithm 1 illustrates this method.

Algorithm 1: Method TB

Input: R^i, M^i
Output: R^i_{TB}
$R^i_{TB} \leftarrow R^i$
for $\tau_{a,b} \in R^i$ **do**
  if $(\exists m \in M^i : v_a \in m) \wedge (\nexists m \in M^i : v_b \in m)$ **then**
    set $v_a \succ v_b$ in R^i_{TB} ;
  else if $(\exists m \in M^i : v_b \in m) \wedge (\nexists m \in M^i : v_a \in m)$ **then**
    set $v_b \succ v_a$ in R^i_{TB} ;
end

For instance, assume that i distributes their points to the six policy options as $C^i = \{30, 40, 10, 20, 0, 0\}$. The multiplication of $VO^i \times {C^i}^T$ returns $U^i = \{100, 90, 80, 80, 70\}$, resulting in $R^i_C : v_1 \succ v_2 \succ v_3 \sim v_4 \succ v_5$. However, if one of the motivations provided by

the participant mentions v_4 and no motivations mention v_3, then the TB method breaks the tie by setting $v_4 \succ v_3$, thus resulting in $R^i_{TB} : v_1 \succ v_2 \succ v_4 \succ v_3 \succ v_5$.

4.1.4 Method MC: Motivations are More Relevant than Choices

There may be an inconsistency between R^i previously estimated for an individual and the values supported by their motivations. That is, R^i indicates $v_b \succ v_a$ but v_a is supported in a motivation $m_j \in M^i$, and v_b is not supported in any motivation. In this case, the MC method prioritizes the value mentioned in the motivation over the one not mentioned, assuming that the value not mentioned is not relevant for individual i in option o_j.

When an inconsistency is detected, we assume that the initial value-option matrix VO^i was inaccurate and update it. In particular, we set the cell of VO^i corresponding to v_b for the option o_j supported by $m_j = \{v_a\}$ to 0. For instance, assume that a participant allocates points to option o_5 where, according to the initial estimate of VO^i (as presented in Table 4), v_1 is relevant but v_4 is not, but mentions v_4 in the motivation. Then, MC adjusts VO^i by setting the cell (v_1, o_5) to 0. We repeat this process for all $v_b : v_b \succ v_a$. Once VO^i is updated for all inconsistencies, we compute the value ranking R^i_{MC} as Algorithm 2 illustrates.

Algorithm 2: Method MC

```
   Input: R^i, M^i, VO^i, V, C^i
   Output: R^i_MC
1  for m_o ∈ M^i do
2  |  for v_a ∈ m_o do
3  |  |  for v_b ∈ V \ {v_a} do
4  |  |  |  if v_a ≺ v_b then
5  |  |  |  |  VO^i(v_b, o) = 0;
6  |  |  end
7  |  end
8  end
9  U^i = VO^i × C^{i^T};
10 R^i_MC = rank(U^i);
```

4.1.5 Method MO: Motivations are Only Relevant for One Policy Option

The motivations M^i provided for different policy options can also bring inconsistencies. For example, consider the initial estimate of VO^i as in Table 4. Further, assume that individual i motivated o_1 with value v_3 ($m_1 = \{v_3\}$), and o_2 with value v_5 ($m_2 = \{v_5\}$). From the notion of valuing as a deliberatively consequential process, from m_1 we can infer that $v_3 \succ v_5$, whereas from m_2 we can infer that $v_5 \succ v_3$.

As in the MC method, when an inconsistency is detected, we assume that the initial value-option matrix VO^i was inaccurate and update it. In particular, we set the cell of VO^i corresponding to the value which is part of the inconsistency but was not mentioned in the provided motivation to 0. From our example, the method would set $VO^i(v_5, o_1)$ and $VO^i(v_3, o_2)$ to 0. Once the VO^i matrix is updated for all the motivations $\times$ options inconsistencies, we compute the value ranking R^i_{MO}. Algorithm 3 illustrates this procedure.

Algorithm 3: Method MO

Input: M^i, VO^i, C^i, V
Output: R^i_{MO}

$VO^i_{MO} \leftarrow VO^i$; /* Temporary copy, we need information from the original VO^i in the next loops */
for $m_a \in M^i : m_a \neq \varnothing$ **do**
  for $m_b \in M^i \setminus \{m_a\}$ **do**
    $V_\alpha = V \setminus \{v : v \in m_a\} : VO^i(v, o_a) == 1$; /* Values supporting o_a in VO^i, except values in m_a */
    for $v_x \in V_\alpha$ **do**
      if $v_x \in m_b$ **then**
        for $v_y \in m_a$ **do**
          $V_\beta = V \setminus \{v : v \in m_b\} : VO^i(v, o_b) == 1$; /* Values supporting o_b in VO^i, except values in m_b */
          if $v_y \in V_\beta$ **then**
            $VO^i_{MO}(v_x, o_a) = 0$;
        end
    end
  end
end
$VO^i \leftarrow VO^i_{MO}$;
$U^i = VO^i \times {C^i}^T$;
$R^i_{MO} = \texttt{rank}(U^i)$;

4.2 Disambiguation Strategy

The disambiguation strategy is intended to drive the interactions between AI agents and participants by addressing the detected inconsistencies between participants' choices and motivations, so as to improve the value preferences estimation process. Inspired by popular AL strategies (Section 2.4), the strategy iteratively targets the participants deemed to be most informative. We associate informativeness with the inconsistency between a participant's choices and motivations, assuming that the largest inconsistencies may reveal the biggest mistakes in the value preferences estimation process. By addressing the most informative participants first, we aim to rapidly improve the quality of value preferences estimation for all participants. Such an approach can improve the estimation of individual value preferences—the more disambiguations are resolved, the more accurately value preferences are expected to be estimated. Furthermore, this approach would also positively impact the downstream application of aggregating value preferences at the population level by decreasing the total computations needed for a more accurate overall value preference estimation.

Figure 4 provides an overview of the proposed strategy. We consider a hybrid participatory setting where the AI agents are equipped with an NLP model tasked to predict the set of values mentioned in each participant's motivations. Then, value preferences are estimated on the basis of the participants' choices and the value labels that are predicted to support each motivation they provide. We propose that AI agents iteratively interact with the participants with the largest detected inconsistencies between the value prefer-

ences estimated from their choices alone and the value preferences estimated from their motivations alone (provided in support of those choices). In our method, the AI agents interact by asking whether the provided motivations have been correctly interpreted (i.e., if the predicted value labels are correct). Other interaction strategies can be implemented (e.g., querying the participants on whether the preference between two values v_a and v_b has been correctly estimated), which we discuss as future work (Section 8).

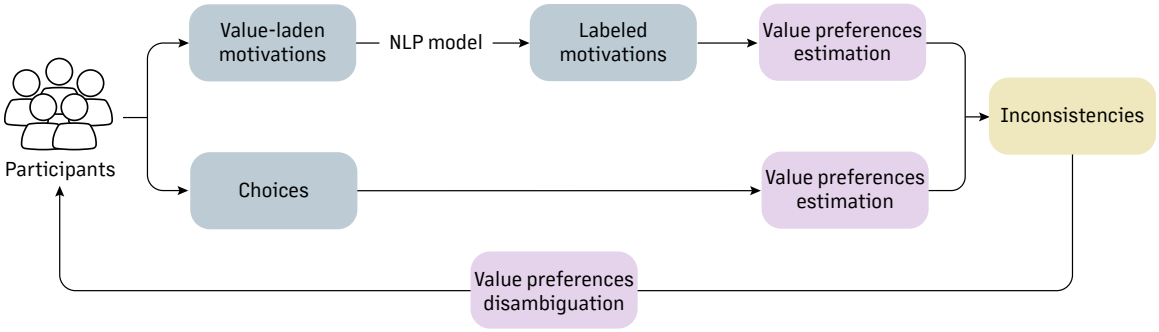


Figure 4: Overview of the proposed disambiguation strategy, guided by the detected inconsistencies between value preferences estimated from participants' choices and motivations.

Our setting is akin to an AL setting where value labels are iteratively retrieved from an annotated dataset (in our experiments, the PVE dataset, as we further elaborate in Section 5.2.2) to train a value classification NLP model. The most informative participants are iteratively selected by the strategy and asked to provide the correct value labels on their motivations, in practice treating the participants themselves as oracles. At every iteration of the AL procedure, we use the current version of the NLP model to predict value labels on all the unlabeled motivations and use the predicted labels to estimate the value preferences of the participants whose motivations are not yet labeled, with both method C and method M. Then, for each participant, we calculate the distance between the value ranking estimated with method C and the value ranking estimated with method M. We use the Kemeny distance (Kemeny & Snell, 1962; Heiser & D'Ambrosio, 2013) to measure the distance between rankings, as it accounts for potential ties between values (Def. 2). The Kemeny distance (d_K) between two value rankings (R^i_C, R^i_M) is defined as:

$$d_K(R^i_C, R^i_M) = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}|x_{(C)jk} - x_{(M)jk}|,$$

where n is the number of objects (in our case, $n = 5$ is the number of values), and $x_{(C)jk}$ is equal to 1 if value j is preferred over value k in ranking R^i_C, equal to -1 in the reverse case, and equal to 0 if the two values are equally preferred. Finally, we choose as the next batch the p participants with the largest Kemeny distance between the value rankings estimated with method C and method M, and retrieve value labels for the motivations they provided. The NLP model is trained with the newly collected annotated motivations, and the AL strategy is re-iterated with the updated version of the NLP model.

5. Experimental Setting

We describe the experiments we perform to evaluate the proposed method[4]. First, we compare the five proposed value preferences estimation methods through a human evaluation procedure. Then, we use the best-performing value preferences estimation method in an AL setting to evaluate the disambiguation strategy by comparing it to traditional AL strategies.

5.1 Value Preferences Estimation

Given the participation process on energy transition using PVE described in Section 3.1, we initialize VO^i by considering a value v_l as relevant for an option o_j if at least t motivations (in our case, we set $t = 20$) among all participants were annotated with v_l for o_j. The resulting initial VO^i matrix (as shown in Table 4) is intended to represent an average rough estimate of the participants' value preferences. We use this as a starting point to apply the methods described in Section 4.1 for all participants. In this way, we create a common starting point to show the effect that the different methods have on tailoring the resulting value rankings to each individual. Nevertheless, other choices for initializing VO^i are equally valid, as we elaborate in Section 7.

		Options					
		o_1	o_2	o_3	o_4	o_5	o_6
Values	v_1	1	1	1	1	1	1
	v_2	1	1	0	1	1	1
	v_3	1	1	1	0	0	0
	v_4	1	1	1	0	0	1
	v_5	1	1	0	0	1	0

Table 4: Initial value-option matrix (VO^i) for the energy transition PVE.

We analyze each method (C, M, TB, MC, and MO) individually, and a sequential combination of the proposed methods in the following order: $MO \Rightarrow MC \Rightarrow TB$. We choose this sequential combination for two reasons: (1) the method TB should be executed last because it does not impact the VO^i matrix directly and thus would not affect the subsequent methods, and (2) we tested both $MO \Rightarrow MC \Rightarrow TB$ and $MC \Rightarrow MO \Rightarrow TB$, and empirically found that the former consistently yields better results (see Appendix A.2 for the comparison). To combine these methods sequentially, we use the ranking and VO^i resulting from MO as input for MC, and the ranking and VO^i resulting from MC as input for TB. Finally, for the individual analysis of the methods TB and MC, that require a previously estimated ranking, we start with the ranking estimated from choices alone (method C). We evaluate these methods based on the resulting value preferences rankings, which we refer to as R_C, R_M, R_{TB}, R_{MC}, R_{MO}, and R_{comb} (where R_{comb} is the result of the sequential combination $MO \Rightarrow MC \Rightarrow TB$).

Evaluation Procedure

Two evaluators, with previous knowledge of values and this specific PVE, were asked to independently judge the value preferences of a subset of participants based on their choices

4. The code is available at `https://github.com/enricoliscio/value-preferences-estimation`

C^i and the provided textual motivations (from which M^i was annotated). We did not describe our value preference estimation methods to the evaluators.

The evaluators were presented with a PVE participant's choices and motivations (but not to the value preferences estimated through our methods), proposed pairs of values (e.g., v_a and v_b), and asked to judge how the two values should be ranked for that participant with the following options: (1) $v_a \succ v_b$; (2) $v_a \prec v_b$; (3) $v_a \sim v_b$; or (4) "I do not know", if they believe there is not enough information to make a proper comparison. Up to four different pairs of values (v_a, v_b) were chosen for each selected PVE participant and judged by the evaluators, with the intent of collecting sufficient information about a participant while increasing the number of analyzed participants.

The values to be compared were randomly selected from a set of value rankings that showed divergence across the methods. Our goal with this procedure is to assess the extent to which the proposed methods estimate value preferences similarly to the human evaluators. Within the envisioned application context described in Section 1, we expect that, as the methods' rankings mirror human intuition, they might provide meaningful feedback to participants in a participatory system.

5.2 Disambiguation Strategy

We test the disambiguation strategy as a sampling strategy in an AL setting, where the motivations' annotations are iteratively retrieved and used to train an NLP model tasked to classify the values that support each motivation. We treat value classification as a multi-label classification task, where each motivation is annotated with zero or more value labels. Since not all provided motivations ought to be value-laden, a motivation may have zero labels in case none of the values in Table 2 is deemed relevant.

5.2.1 Model Selection

Multi-label BERT (Devlin et al., 2019) has been shown to produce state-of-the-art performances on similar value classification tasks at the time of writing (Liscio et al., 2022; Kiesel et al., 2022; Huang et al., 2022; Qiu et al., 2022). As the PVE corpus was originally collected in Dutch, we chose to employ RobBERT (Delobelle et al., 2020), a BERT variant considered state-of-the-art for the Dutch language at the time of writing. However, due to the more widespread usage of the English language in NLP models, we also decided to translate the corpus to English and test two models trained in English—a RoBERTa model (Liu et al., 2019) (similar to the Dutch model) and a comparably sized model with a different architecture, XLNet (Yang et al., 2019). We further detail our experiments and hyperparameters search in Appendix A.3, with Table 5 showing the performances with the best resulting models. As noticeable, the difference between the three tested models is minimal. Thus, we opted for the RobBERT Dutch model to employ the original data.

	RobBERT (Dutch)	RoBERTa (English)	XLNet (English)
micro F_1-score	0.64	0.65	0.65
macro F_1-score	0.63	0.64	0.64

Table 5: Micro and macro F_1-scores with the tested Dutch and English models.

5.2.2 Training Procedure

In a typical AL setting, a large pool of unlabeled data points is initially available. A sampling strategy is used to iteratively select a set of unlabeled data points that are to be annotated and added to the pool of labeled data points (i.e., the data points that are used to train the NLP model). Together, labeled and unlabeled data points constitute the set of data points that are available for training. In addition, a set of labeled test data points (which are not available to be selected through the sampling strategy) is kept aside to evaluate the NLP model. In our case, we employ the annotated PVE dataset described in Section 3.1 to simulate the AL procedure. That is, we initially set aside a test set, and pretend that no labels are available for the remaining data points (which constitute the initial set of unlabeled data points). As the sampling strategy selects the unlabeled data points to be annotated, we retrieve the corresponding annotations from our PVE dataset and add these labeled data points to the set of labeled data points.

At every AL iteration, we have a set of labeled motivations (whose labels have been retrieved and that are used to train the NLP model), a set of unlabeled motivations (whose labels can be retrieved if selected by the sampling strategy), and a set of test motivations (that are only used for evaluation). Analogously, we refer to the PVE participants who wrote the motivations in the corresponding sets as labeled participants, unlabeled participants, and test participants. At every iteration, the model is trained with the labeled motivations, and used to predict labels on the unlabeled motivations. With the predicted labels, the value preferences of the unlabeled participants are estimated. The disambiguation strategy is then used to select the p unlabeled participants with the most inconsistent value preferences estimated from choices and from motivations alone. The p participants are added to the set of labeled participants, the labels of the motivations provided by the participants are retrieved, and the motivations are added to the set of labeled motivations.

As is common in AL settings, we warm up the NLP model by initializing the set of labeled participants with 10% of the available participants, and the set of labeled motivations with the motivations provided by those participants. At each iteration, we train the NLP model with the labeled motivations. We use the trained model to predict labels on the test motivations and use these labels to (1) estimate the value preferences of the test participants with the best-performing value preferences estimation method, and (2) evaluate the performance of the NLP model. We then use the disambiguation strategy to select $p = 39$ participants, so as to add 5% of the available participants to the labeled participants set at each iteration. We iterate the procedure for 5 iteration steps and repeat it in a 10-fold cross-validation.

5.2.3 Evaluation Procedure

We evaluate how the proposed disambiguation strategy drives the NLP model performance and the estimation of value preferences, comparing it to the respective toplines and baselines.

We perform 10-fold cross-validation to measure the performance of the NLP model trained on all available data and use the result as the NLP topline during the AL procedure. We use a model trained on all data to predict labels on all the motivations and use the predicted labels to estimate all participants' value preferences with the best-performing value preferences estimation method. We treat the resulting value rankings as value prefer-

ences topline during the AL procedure, as they represent the best possible value rankings that can be estimated with the mistakes introduced by using the labels predicted by an NLP model instead of the ground truth annotations. At every iteration of the AL procedure, we compare the NLP performance on the test set to the NLP topline, and the estimated value preferences of the test participants to the value preferences topline. For the NLP performance, we report the micro F_1-score as it accounts for the label distribution (which is imbalanced, see Table 3). Finally, for the value preferences estimation performance, we report the Kemeny distance between the estimated value preferences of the test participants and the corresponding value preferences topline.

We compare the results to two baselines. First, we employ the uncertainty sampling strategy (Section 2.4) to select 5% of motivations (i.e., 145 motivations) at each iteration, similarly to the evaluated disambiguation strategy. This strategy is solely driven by motivations informativeness, ignoring the connection between the motivations and their authors. We choose this strategy as a baseline since traditional NLP AL strategies are solely driven by information about the NLP task, as described in Section 2.4. Second, we employ a random baseline, where at each iteration 5% random participants ($p = 39$ participants, similarly to the proposed disambiguation strategy) and their motivations are added to the labeled set. With both our proposed strategy and the baselines, we plot the trend of the NLP and value preferences estimation performances throughout the progressive iterations. We compare them with each other and to the corresponding toplines.

6. Results and Discussion

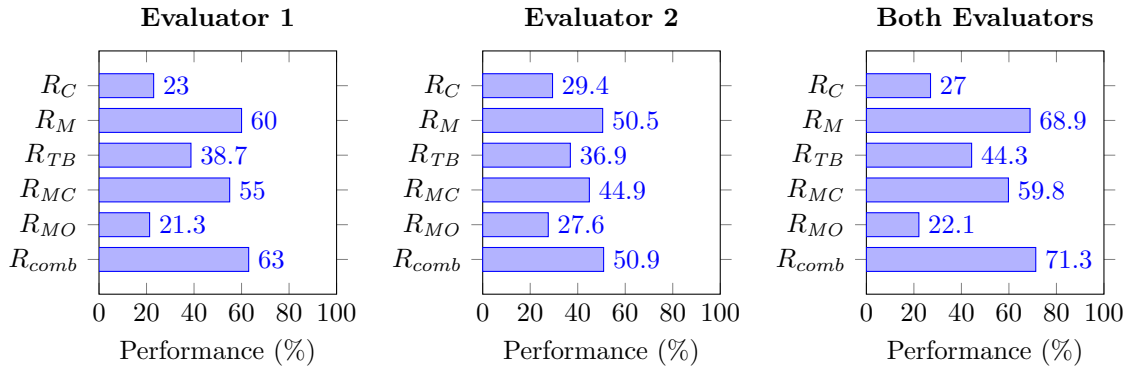
We present and discuss the evaluation of the value preferences estimation methods and the interactive disambiguation strategy.

6.1 Value Preferences Estimation

When comparing the five proposed value preferences estimation methods, we aim to answer two questions: (1) How well can each method estimate value preferences compared to humans? (2) How does the estimation of value preferences differ among the methods proposed?

The evaluators performed 1047 comparisons. We discard the responses indicating that there was not enough information to judge values preference ("I do not know"), reducing the analyzed set to 766 total responses by either one of the evaluators. Figures 5a and 5b present the performance of each method in terms of matching each evaluator's responses. These comparisons overlapped 269 times (i.e., the annotators performed the same comparisons). Considering this subset of overlapping comparisons, we find an agreement in 122 (45.35%) and disagreement in 147 (54.65%) comparisons, resulting in a Kappa score of 0.247, which is considered a fair agreement (Landis & Koch, 1977). To mitigate the effect of individual biases, in the remainder of the analysis, we focus on the pairwise comparisons that evaluators agreed on, as presented in Figure 5c.

As Figure 5 displays, the rankings R_M, R_{MC}, and R_{comb} provide the best performance in terms of human-like value preferences estimation. When compared to R_C, the combined method R_{comb} estimated value preferences 2.64 times more similarly to humans (considering the subset where evaluators agreed). Further, we observe that R_M and R_{MC} also performs



(a) Overlap with Evaluator 1. (b) Overlap with Evaluator 2. (c) Overlap where they agree.

Figure 5: Performance of the value preferences estimation methods, measured as the overlap with the evaluators' answers.

better than R_C. The only exception in terms of performance is R_{MO}, which performs slightly worse than R_C. These findings show that combining choices and motivations in estimating value preferences can significantly increase the degree to which an automated method can estimate value preferences similarly to humans, with respect to using only choices.

Finally, we notice that the performance of R_M is similar to the performance of R_{comb}. This is to be expected, as R_{comb} prioritizes motivations over choices, and R_M only employs motivations to estimate value preferences. The visibly better performance of R_M with respect to R_C further motivates the need to consider textual motivations to estimate value preferences that are consistent with human evaluation. With our dataset, combining choices and motivations led to slightly better results than employing just the motivations. Further experiments with other data are needed to confirm this observation.

6.1.1 Comparative Analysis

For each method, we average the value preference rankings (that is, the position that the values have in the ranking that results after applying the method). We indicate with $\succ$ the values that have significantly different average rankings ($p \leq 0.05$) and with $\succeq$ the values that do not have significantly different averages. The following are the resulting average rankings per each different method:

- R_C: $v_1 \succ v_2 \succ v_4 \succ v_5 \succeq v_3$
- R_M: $v_3 \succ v_1 \succeq v_2 \succ v_5 \succeq v_4$
- R_{TB}: $v_1 \succ v_2 \succ v_4 \succ v_5 \succ v_3$
- R_{MC}: $v_1 \succ v_2 \succ v_5 \succ v_3 \succ v_4$
- R_{MO}: $v_1 \succ v_2 \succ v_4 \succ v_5 \succeq v_3$
- R_{comb}: $v_1 \succeq v_2 \succ v_5 \succeq v_3 \succeq v_4$

Method C ranked the value v_1 as the most important for all individuals, regardless of their choices, due to the characteristics of the initial value option-matrix (VO^i) in Table 4, which considers v_1 relevant for all choice options. As we attribute the minimum ordinal ranking for the values in case of ties (Def. 2), any choices would lead to R^i_C with v_1 as (one of) the most important value(s), except for method M which does not consider choices.

Let R_C be a baseline for comparison. Figure 6 indicates how many positions the final ranking changed across values (we do not consider method M since it did not use R_C as baseline). For example, consider two rankings $R_1 : v_1 \succ v_2 \succ v_3 \succ v_4 \succ v_5$ and $R_2 : v_2 \succ v_3 \succ v_1 \succ v_4 \succ v_5$. We consider four position changes from R_1 to R_2: v_1 changed from the first to the third position (two changes), v_2 changed from the second to the first position (one change), and v_3 changed from the third to the second position (one change).

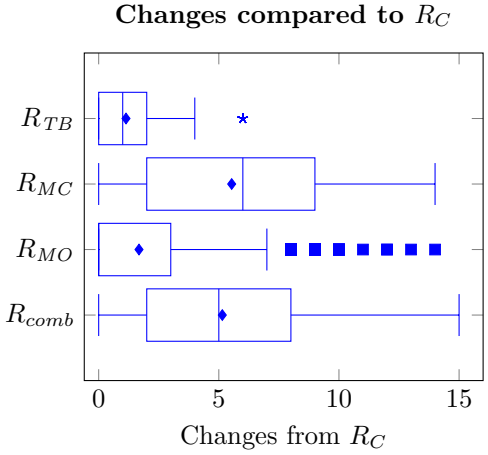


Figure 6: Average changes in the value rankings when compared to R_C.

Rankings R_{TB} and R_{MO} barely deviate from the average R_C. Instead, R_{MC} and the combined approach R_{comb} show significant deviation from R_C, indicating a larger difference at an individual value preferences level. The large deviation and the good performance (see Figure 5) of these two methods suggest that they estimate individually tailored value preferences that align with human intuition.

6.2 Disambiguation Strategy

First, we report the results of the toplines. The NLP topline resulted in an average micro F_1-score of 0.64 (Table 5), which is slightly lower than similar value classification tasks (Liscio et al., 2022; Huang et al., 2022), likely due to the smaller dataset size. For the value preferences topline, we use the predicted motivation labels to estimate value rankings through the R_{comb} method (the best-performing value preferences estimation method). The value preferences topline resulted in an average Kemeny distance of 1.88 (with 2.88 standard deviation) from the value rankings estimated with the $MO \Rightarrow MC \Rightarrow TB$ method (with the resulting ranking R_{comb}) by using the ground truth annotations on the motivations. We use these toplines to measure the trend of the results throughout the AL iterations.

We report the results of our experiments in Figures 7 and 8. In all experiments, at every iteration we used the tested strategy to select 5% of the data to be added to the set of labeled data. However, since different participants provided different numbers of motivations, selecting the motivations provided by 5% of the participants may not correspond to 5% of all available motivations. In Figures 7 and 8, we show on the x-axis the number of motivations used for training the NLP model at the corresponding iteration. While

that corresponds to exactly 5% increments in the case of the uncertainty strategy (which selects 5% of the motivations at every iteration), it is not the case for the random and disambiguation strategies (which selects 5% of the participants at every iteration).

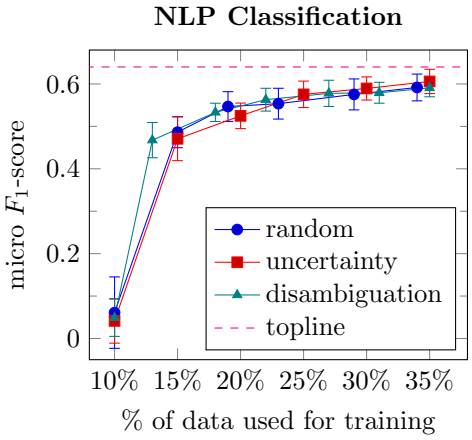


Figure 7: NLP performance (micro F_1-score), compared to the NLP topline (dashed horizontal line).

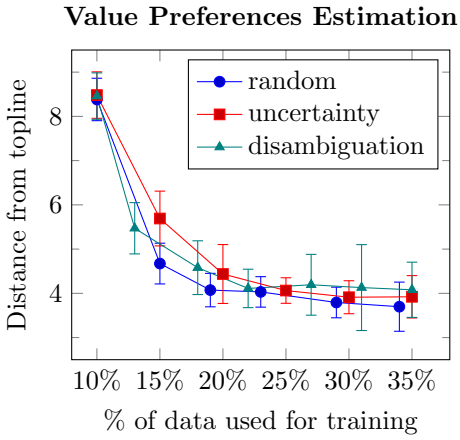


Figure 8: Value preferences estimation performance, measured as average Kemeny distance from the value preferences topline.

The random strategy has a varying step size that roughly averages to 5%, as expected by a strategy that randomly selects participants. Instead, the step size of the disambiguation strategy is consistently smaller than the other two (for this strategy we plot six steps, as opposed to five for the other strategies), meaning that at every iteration the strategy chooses participants who have provided less motivations than the average participant. This empirically matches the intuition behind the strategy—participants who have provided few motivations have a R_M (value ranking calculated from motivations alone) that is mostly composed of ties between values. Such undetermined R_M have a large distance from the corresponding R_C, which instead considers all the choices provided by the participants.

The NLP performances of the model trained with the disambiguation strategy and with the two baseline strategies (uncertainty and random) are illustrated in Figure 7. No significant difference between the compared methods is visible, as all three strategies lead to a rapid improvement in performances that approaches the NLP topline when roughly 30% of the available motivations are used for training. In line with these results, experimental findings (Ein-Dor et al., 2020) show that there is no single AL strategy that outperforms all others across different datasets, and, in some cases, no significant difference is observable with the random strategy. Ultimately, these results demonstrate that the proposed disambiguation strategy, despite being guided by the downstream task of estimating value preferences, does not significantly affect the NLP performance.

Figure 8 presents the value preferences estimation performance of the three compared strategies, measured as the average Kemeny distance between the value preferences estimated with the labels predicted by the current iteration of the NLP model and the value preferences topline. First, we remind that the topline has been calculated with the label predictions resulting from a model trained with all available data. However, the training

process with the tested strategies is performed as 10-fold validation, thus a different subset of the dataset is used for training in each fold. Consequently, we do not expect the Kemeny distance to approach zero, as different data was used during the training process (thus resulting in different individual value preferences). Still, the topline reference allows us to compare the value preferences estimation performance trend of the three strategies.

We observe that the value preferences estimation performance trend is similar for all three strategies, leading to a rapid decrease in distance from the topline that mirrors the rapid improvement in the F_1-score. While the results are comparable when 20% or more motivations are used for training, the results with $\sim$15% of the training data show small differences—while the F_1-score performances at this stage are almost identical, there is a small difference in value preferences estimations. In particular, the uncertainty strategy (which ignores the link between users and motivations) is worse than the other two tested strategies, which motivates the usage of a user-driven strategy instead of a motivation-driven strategy. However, the differences are not sufficiently large to draw a definitive conclusion.

Overall, we notice no significant difference between the proposed strategy and the baselines. We discuss two possible reasons. First, the NLP performance is the biggest driver of value preferences estimation performance—in practice, the more motivations are correctly labeled, the more accurate the value preferences estimation is. With the analyzed data and the relatively small dimension of the dataset, no significant difference is noticeable between the tested strategies in NLP performance, including the random strategy, resulting in a similar trend in the value preferences estimation performance. Second, the distance between R_M and R_C may not be the best indicator for the informativeness of a user. Considering the annotations from Section 3.1, there is a distance of 8.0 (with 3.5 standard deviation) between R_M and R_C estimated for the same users. Thus, large distances between the two rankings may not be particularly informative in this dataset. However, we believe that a strategy driven by the downstream application may be particularly useful in similar settings, as we elaborate as Future Work.

7. Limitations

We discuss the main limitations of our experimental results.

First, we discuss the generalizability of our experimental setting. As described in Section 1, we envision our value estimation and disambiguation methods to be employed during an ongoing deliberation. However, to showcase the proposed methods, we tested them on a concluded survey and used third-party annotations to evaluate them, thus limiting the validity of our results due to the subjective nature of the annotation and evaluation task. Such limitations can be addressed, for example, by asking annotators to perform vicarious annotations (Weerasooriya et al., 2023) or by turning to a diverse sample of annotators (van der Meer et al., 2024). However, the most effective approach would be to consult the participants themselves, as we further elaborate in Section 8.

Second, we discuss the initialization of VO^i. We decided to initialize all initial VO^i's identically (as motivated in Section 5.1) to demonstrate the effectiveness of the value preferences estimation strategies in tailoring value preferences to the individuals. However, in different settings, other initialization strategies could be equally valid. For instance, VO^i could be initialized based on (1) the results of a previous (or another) session of deliberation,

(2) a self-reported estimate of value preferences from the participants, (3) an initial estimate from the policy-makers, or (4) a demographic grouping of initially estimated preferences. The effectiveness of our proposed value preferences estimation methods ought to be studied with different initializations of VO^i. Methods *MO* and *MC* adjust VO^i by turning ones into zeros—precisely, the best-performing method *comb* leads the average VO^i matrix from having 21 to having 15.71 ± 3.74 ones. Despite resulting in still populated VO^i matrices in our experiments, this effect could be detrimental with a less populated initial VO^i. Similarly, different strategies (or applications thereof) should be devised for repeated use over several deliberation sessions, both to avoid matrix sparsity and that the latest rounds of deliberation (accidentally) override the results obtained in the previous.

Third, we discuss the choice of the value labels that we employ in our experiments. As described in Section 3.1, we perform our experiments with a subselection of the values identified by Kaptein (2020). We choose so as our experiments are not intended as a comprehensive analysis of the value preferences of the survey participants, but rather a simplified scenario to showcase our methods. The application of our methods in a real deliberation setting ought to be able to handle (1) larger lists of values (to which our methods are compatible), and (2) changing lists of values, which may be iteratively updated during the deliberation with methods such as the one proposed by Liscio et al. (2022). Furthermore, in the dataset we used, values were annotated only when supporting the motivations and thus the related choice. However, choices could also demote values, and as so be reflected in the related motivations. Recent works have investigated this approach to value valence (i.e., that values could be promoted or demoted by actions) for value preferences aggregation (Lera-Leri et al., 2024) and value classification in text (Sorensen et al., 2024). In this case, our methods ought to be updated. First, the NLP model should be trained to predict a value label behind the motivations that range from positive to negative. Next, the notion of valence could be inserted into the disambiguation strategy (e.g., by prioritizing participants with motivations with opposite valences for the same values) and preferences estimation strategy (e.g., by accounting for the valence of the values when addressing inconsistencies).

Finally, we discuss the validity of our machine learning experiments. We experimented on a (relatively small) dataset composed of survey answers in Dutch. Further experiments are needed to validate our findings with other types of data (e.g., conversational) and underrepresented languages, and to validate the impact of label distribution on both the results with the NLP model and disambiguation strategy. Furthermore, the proposed disambiguation strategy is sensitive to outlier participants, since it targets the largest inconsistencies between participants' choices and motivations. This creates the risk that the NLP model—and, consequently, the value preferences estimation results—are built on data that is not representative of the overall population, or worse, on noisy data. While the distinction between noisy and minority voices in subjective tasks is under debate (Plank, 2022; Cabitza et al., 2023), our envisioned application addresses the problem by consulting the participants themselves, as we further elaborate in Section 8.

8. Conclusion and Future Directions

We introduce a method for estimating how participants prioritize competing values in a hybrid participatory system, through a disambiguation strategy aimed at guiding the in-

teractions between AI agents and participants. Our method directly targets the detected inconsistencies between participants' choices and motivations.

First, we propose and compare methods for an AI agent to estimate the value preferences of individuals from one's choices and value-laden motivations, with the goal of generating an ordered value ranking within the analyzed context. We aim to improve the estimation of value preferences by prioritizing value preferences estimated from motivations over value preferences estimated from choices alone. We test our methods in the context of a large-scale survey on energy transition. Through a human evaluation, we show that incorporating motivations to deal with conflicts in value preferences improves the performance of value preferences estimation by more than two times (in terms of similarity to human evaluators' value preferences estimation) and yields preferences that are more individually tailored.

Second, we propose a disambiguation strategy to drive the interactions between AI agents and participants, with the intent of improving the value preferences estimation performance. Our strategy prioritizes the interaction with the participants whose value preferences estimated from choices alone are most different from the value preferences estimated from motivations alone, following the rationale that such participants would be the most informative for rapidly adjusting and improving the value preferences estimation process. However, our results show no significant difference with compared baseline strategies, including a strategy where interactions with users are randomly determined.

Despite the inconclusive results, we believe that our proposed disambiguation strategy opens novel research avenues. Such a hybrid approach to an interaction strategy for value preferences disambiguation can help not only iteratively address algorithmic mistakes, but also foster self-reflection in participants by situating their estimated value preferences in specific contexts and choices (Liscio et al., 2023), which has been shown to raise awareness and lead to changing perspectives (Lim et al., 2019). A strategy driven by the downstream task of value preferences estimation helps in integrating the different components involved in the value preferences estimation process (value label classification and aggregation of one's choices and motivations). Further, different disambiguation approaches could be tested. For instance, the strategy could target the participants with the most different choice distribution from the average, or with the largest amount of ties in their estimated value rankings.

We identify additional directions for future work. On the one hand, we suggest exploring other approaches to associate values with choice options beyond a binary matrix—for instance, given n considered values, by using an $n \times n$ matrix that reflects pairwise comparisons among all values, thus allowing non-transitive value preferences (Alós-Ferrer et al., 2022). On the other hand, using our proposed methods during an ongoing deliberation would open additional future work avenues. In such a setting, we envision a language model to be trained to recognize values in text through the disambiguation strategy, use it to classify the values in the motivations, and use this information to estimate value preferences. An interesting extension compatible with the proposed methods is to let participants themselves provide direct feedback to the AI agent, instead of relying on external evaluators. Additionally, following the self-reflection fostered by the disambiguation strategy, participants may be offered the option to adjust their choices or the estimated value preferences directly, instead of being limited to providing the correct value label supporting

their motivations. Machine learning methods could then be employed for value preferences estimation, learning directly from the feedback provided by the participants.

Our work has the potential to contribute to value alignment between AI and humans. The estimated value preferences can serve as a starting point for the operationalization of values, e.g., for the synthesis of value-aligned normative systems (Serramia et al., 2021; Montes & Sierra, 2022), as a foundation for international regulatory systems (Bajgar & Horenovsky, 2023), or to formulate ethical principles through a combination of machine learning and logic (Kim et al., 2021). In the context of a hybrid participatory system, the estimated individual value preferences can be aggregated at a societal level (Lera-Leri et al., 2024) to provide policy-makers with an overview of the value preferences of a population.

Acknowledgments

Enrico Liscio and Luciano C. Siebert contributed equally to this work. This work was partially supported by TU Delft's AiTech initiative, by TAILOR, a project funded by the EU Horizon 2020 research and innovation programme under GA No 952215, and by the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022). We thank Lionel Kaptein, Shannon Spruit, and Jeroen van den Hoven for their help with previous iterations of this work.

References

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., & Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, *53*(8), 18–28.

Alós-Ferrer, C., Fehr, E., & Garagnani, M. (2022). Identifying nontransitive preferences. Working paper 415, University of Zurich, Department of Economics, Zurich.

Alshomary, M., Baff, R. E., Gurcke, T., & Wachsmuth, H. (2022). The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '22, pp. 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Araque, O., Gatti, L., & Kalimeri, K. (2020). MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowledge-Based Systems*, *191*, 1–29.

Araque, O., Gatti, L., & Kalimeri, K. (2021). The Language of Liberty: A preliminary study. In *Companion Proceedings of the Web Conference 2021*, WWW '21 Companion, pp. 1–4, Ljubljana, Slovenia. Association for Computing Machinery.

Bahgat, M., Wilson, S. R., & Magdy, W. (2020). Towards Using Word Embedding Vector Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 919–923, Atlanta, Georgia. AAAI Press.

Bajgar, O., & Horenovsky, J. (2023). Negative Human Rights as a Basis for Long-term AI Safety and Regulation. *Journal of Artificial Intelligence Research*, *76*, 1043–1075.

Benabou, R., Falk, A., Henkel, L., & Tirole, J. (2020). Eliciting Moral Preferences: Theory and Experiment. Tech. rep., Princeton University.

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In *Proceedings of the 9th International Conference on Web and Social Media*, pp. 31–40.

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 6860–6868.

Dallhammer, E., Gaugitsch, R., Neugebauer, W., & Böhme, K. (2018). Spatial planning and governance within eu policies and legislation and their relevance to the new urban agenda. Tech. rep., European Committee of the Regions: Bruxelles, Belgium.

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3255–3265, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '19, p. 4171–4186.

Dietz, T., & Stern, P. C. (1995). Toward a theory of choice: Socially embedded preference construction. *Journal of Socio-Economics*, *24*(2), 261–279.

Dignum, V. (2017). Responsible Autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI '17, pp. 4698–4704.

Eckersley, P. (2019). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). In *CEUR Workshop Proceedings*.

Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., & Slonim, N. (2020). Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pp. 7949–7962, Online. ACL.

Erdmann, A., Wrisley, D. J., Allen, B., Brown, C., Cohen-Bodénès, S., Elsner, M., Feng, Y., Joseph, B., Joyeux-Prunel, B., & de Marneffe, M. C. (2019). Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '19, pp. 2223–2234, Minneapolis, Minnesota, USA. ACL.

Franco, C., & Ghisetti, C. (2022). What shapes the "value-action" gap? The role of time perception reconsidered. *Economia Politica*, *39*(3), 1023–1053.

Frankfurt, H. (2018). Freedom of the will and the concept of a person. In *Agency And Responsiblity*, pp. 77–91. Routledge.

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press.

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, *50*(1), 344–361.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, Vol. 47, pp. 55–130. Elsevier, Amsterdam, the Netherlands.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Haag, D. (2019). National Climate Agreement - The Netherlands. Tech. rep., Dutch Ministry of Economic Affairs and Climate.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Hämäläinen, N. (2016). *Descriptive Ethics: What does moral philosophy know about morality?* Springer.

Heath, H., & Cowley, S. (2004). Developing a grounded theory approach: a comparison of Glaser and Strauss. *International Journal of Nursing Studies*, *41*(2), 141–150.

Heiser, W. J., & D'Ambrosio, A. (2013). Clustering and prediction of rankings within a kemeny distance framework. In *Algorithms from and for Nature and Life*, pp. 19–31. Springer International Publishing, Cham.

Hill, P. L., & Lapsley, D. K. (2009). Persons and situations in the moral domain. *Journal of Research in Personality*, *43*(2), 245–246.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, *53*, 232–246.

Huang, X., Wormley, A., & Cohen, A. (2022). Learning to Adapt Domain Shifts of Moral Values via Instance Weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22)*, pp. 121–131, Barcelona, Spain. ACM.

Iandoli, L., Quinto, I., De Liddo, A., & Buckingham Shum, S. (2016). On online collaboration and construction of shared knowledge: Assessing mediation capability in computer supported argument visualization tools. *Journal of the Association for Information Science and Technology*, *67*(5), 1052–1067.

Itten, A., & Mouter, N. (2022). When digital mass participation meets citizen deliberation: combining mini-and maxi-publics in climate policy-making. *Sustainability*, *14*(8), 4656.

Kaptein, L. (2020). Participatory value evaluation as a tool for value extraction and opinion mining: Reduce manual data analysis by automated value extraction. Master's thesis, Delft University of Technology.

Kemeny, J. G., & Snell, L. J. (1962). Preference ranking: an axiomatic approach. In *Mathematical models in the social sciences*, pp. 9–23. Ginn New York.

Kenter, J. O., Reed, M. S., & Fazey, I. (2016). The deliberative value formation model. *Ecosystem Services*, *21*, 194–207.

Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., & Stein, B. (2022). Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '22, pp. 4459–4471, Dublin, Ireland. ACL.

Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, *70*, 871–890.

Lafont, C. (2015). Deliberation, participation, and democratic legitimacy: Should deliberative mini-publics shape public policy?. *Journal of political philosophy*, *23*(1), 40–63.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159.

Le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009). Values as Lived Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1141–1150. ACM Press.

Lera-Leri, R. X., Liscio, E., Bistaffa, F., Jonker, C. M., Lopez-Sanchez, M., Murukannaiah, P. K., Rodriguez-Aguilar, J. A., & Salas-Molina, F. (2024). Aggregating value systems for decision support. *Knowledge-Based Systems*, *287*, 111453.

Liao, Q. V., & Muller, M. (2019). Enabling Value Sensitive AI Systems through Participatory Design Fictions..

Lim, C. Y., Berry, A. B., Hartzler, A. L., Hirsch, T., Carrell, D. S., Bermet, Z. A., & Ralston, J. D. (2019). Facilitating Self-reflection about Values and Self-care among Individuals with Chronic Conditions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 12, Glasgow, UK. ACM.

Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C. M., Kalimeri, K., & Murukannaiah, P. K. (2023). What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, ACL '23, pp. 14113–14132, Toronto, Canada. ACL.

Liscio, E., Dondera, A. E., Geadau, A., Jonker, C. M., & Murukannaiah, P. K. (2022). Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2727–2745, Seattle, WA, USA. ACL.

Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R. I., Jonker, C. M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., & Murukannaiah, P. K. (2023). Value Inference in Sociotechnical Systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pp. 1774–1780, London, United Kingdom. IFAAMAS.

Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., & Murukannaiah, P. K. (2022). What Values Should an Agent Align With? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, *36*(23), 32.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach..

Miller, S. (2016). *Design for Values in Institutions*, pp. 1–11. Springer Netherlands, Dordrecht.

Mindermann, S., & Armstrong, S. (2018). Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*, NeurIPS '18, pp. 5598–5609, Montreal, Canada. Curran Associates, Inc.

Montes, N., & Sierra, C. (2022). Synthesis and Properties of Optimally Value-Aligned Normative Systems. *Journal of Artificial Intelligence Research*, *74*, 1739–1774.

Mouter, N., Hernandez, J. I., & Itten, A. V. (2021). Public participation in crisis policymaking. How 30, 000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures. *PLoS ONE*, *16*(5), 1–42.

Ng, A., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, p. 663–670, Stanford, CA, USA. Cambridge University Press.

Park, J., Liscio, E., & Murukannaiah, P. (2024). Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 654–673, St. Julian's, Malta. ACL.

Pavan, M. C., Santos, V. G., Lan, A. G. J., Martins, J., Santos, W. R., Deutsch, C., Costa, P. B., Hsieh, F. C., & Paraboni, I. (2020). Morality Classification in Natural Language Text. *IEEE Transactions on Affective Computing*, *3045*(c).

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC. *Mahway: Lawrence Erlbaum Associates*, *71*.

Pigmans, K., Aldewereld, H., Dignum, V., & Doorn, N. (2019). The Role of Value Deliberation to Improve Stakeholder Participation in Issues of Water Governance. *Water Resources Management*, *33*(12), 4067–4085.

Pini, M. S., Rossi, F., Venable, K. B., & Walsh, T. (2005). Aggregating partially ordered preferences: impossibility and possibility results. In *Proceedings of the 10th conference on Theoretical aspects of rationality and knowledge*, pp. 193–206.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pommeranz, A., Detweiler, C., Wiggers, P., & Jonker, C. M. (2012). Elicitation of situated values: Need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology*, *14*(4), 285–303.

Ponizovskiy, V., Ardag, M., Grigoryan, L., Boyd, R., Dobewall, H., & Holtz, P. (2020). Development and Validation of the Personal Values Dictionary: A Theory-Driven Tool for Investigating References to Basic Human Values in Text. *European Journal of Personality*, *34*(5), 885–902.

Qiu, L., Zhao, Y., Li, J., Lu, P., Peng, B., Gao, J., & Zhu, S.-C. (2022). ValueNet: A New Dataset for Human Value Driven Dialogue System. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, AAAI '22, pp. 11183–11191.

Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2021). A Survey of Deep Active Learning. *ACM Computing Surveys*, *54*(9), 1–40.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, *36*(4), 105–114.

Russell, S. J. (2019). *Human Compatible: AI and the Problem of Control* (1st edition). Viking Press.

Scheffler, S. (2012). Valuing. In *Equality and Tradition: Questions of Value in Moral and Political Theory* (1st edition)., chap. 1, pp. 15–40. Oxford University Press.

Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, *2*(1), 1–20.

Serramia, M., Lopez-Sanchez, M., Moretti, S., & Rodriguez-Aguilar, J. A. (2021). On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems*, *35*(42).

Settles, B. (2012). *Active Learning*. Morgan & Claypool.

Shortall, R., Itten, A., van der Meer, M., Murukannaiah, P. K., & Jonker, C. M. (2022). Reason against the machine? future directions for mass online deliberation. *Frontiers in Political Science*, *4*, 1–17.

Siebert, L. C., Liscio, E., Murukannaiah, P. K., Kaptein, L., Spruit, S. L., van den Hoven, J., & Jonker, C. M. (2022). Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*, pp. 114–127, Amsterdam, the Netherlands. IOS Press.

Smith, M., Lewis, D., & Johnston, M. (1989). Dispositional theories of value. *Aristotelian Society Supplementary Volume*, *63*(1), 89–174.

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., & Choi, Y. (2024). Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 19937–19947.

van der Meer, M., Falk, N., Murukannaiah, P. K., & Liscio, E. (2024). Annotator-Centric Active Learning for Subjective NLP Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18537–18555, Miami, FL, USA. ACL.

van der Meer, M., Vossen, P., Jonker, C., & Murukannaiah, P. (2023). Do Differences in Values Influence Disagreements in Online Discussions?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pp. 15986–16008, Singapore. ACL.

Weerasooriya, T., Dutta, S., Ranasinghe, T., Zampieri, M., Homan, C., & KhudaBukhsh, A. (2023). Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11648–11668, Singapore. Association for Computational Linguistics.

Wilson, K. G., Sandoz, E. K., Kitchens, J., & Roberts, M. (2010). The valued living questionnaire: Defining and measuring valued action within a behavioral framework. *Psychological Record*, *60*(2), 249–272.

Wilson, S. R., Shen, Y., & Mihalcea, R. (2018). Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *Proceedings of the 10th International Conference on Social Informatics*, SocInfo '18, pp. 455–470, St. Petersburg, Russia. Springer.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, NeurIPS '19, pp. 5754–5764, Vancouver, BC, Canada.

Zhang, Y., Lease, M., & Wallace, B. C. (2017). Active Discriminative Text Representation Learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 3386–3392, San Francisco, California, USA.

Zhang, Z., Strubell, E., & Hovy, E. (2022). A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, pp. 6166–6190. ACL.

Zhao, Y., Zhang, H., Zhou, S., & Zhang, Z. (2020). Active Learning Approaches to Enhancing Neural Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1796–1806, Online. ACL.

Zintgraf, L. M., Roijers, D. M., Linders, S., Jonker, C. M., & Nowé, A. (2018). Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, Vol. 2, pp. 1477–1485.