Estimating Value Preferences in a Hybrid Participatory System

Luciano C. SIEBERT [a,1], Enrico LISCIO [a], Pradeep K. MURUKANNAIAH [a], Lionel KAPTEIN [a], Shannon SPRUIT [b], Jeroen van den HOVEN [a] and Catholijn JONKER [a],

[a] *Delft University of Technology, Delft, The Netherlands*
[b] *Populytics, Leiden, The Netherlands*

Abstract. We propose methods for an AI agent to estimate the value preferences of individuals in a hybrid participatory system, considering a setting where participants make choices and provide textual motivations for those choices. We focus on situations where there is a conflict between participants' choices and motivations, and operationalize the philosophical stance that "valuing is deliberatively consequential." That is, if a user's choice is based on a deliberation of value preferences, the value preferences can be observed in the motivation the user provides for the choice. Thus, we prioritize the value preferences estimated from motivations over the value preferences estimated from choices alone. 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 proposed methods can be integrated in a hybrid participatory system, where artificial agents ought to estimate humans' value preferences to pursue value alignment.

Keywords. Value alignment, participatory systems, responsible AI

1. Introduction

Enhancing citizen participation is high on the European policy agenda [1]. Initiatives to foster citizens' political power and engagement have been proposed through the use of digital platforms for participatory decision making [2,3] and deliberation [4,5]. Not only governments, but all organizations striving for user-centric decision making must engage the stakeholders (e.g., user, consumer, citizen) in the decision-making processes and actively involve them in all important decisions. In this context, participatory systems need to not only provide a mechanism for preference elicitation on contextual choices, but also align at a deeper level on the individuals' democratic values.

Values serve as the standards or criteria to justify one's opinions and actions, and are intrinsically linked to goals [6]. Values form an ordered system of priorities and it is the relative importance one gives to values (i.e., one's *value preference*) that guides action. However, how individuals ascribe relative priorities among values can vary significantly for each person, socio-cultural environment [7], and context [8,9].

[1] Corresponding Author: L.CavalcanteSiebert@tudelft.nl

Artificial Intelligence (AI) techniques can enable mass participation in deliberative processes. However, the misuse of AI can lead to ethical impacts that undermine the expected benefits of such systems. Avoiding these impacts requires properly estimating human value preferences, representing the values in a context-sensitive manner, translating the values into technical requirements, creating means to deal with moral dilemmas, and verifying value alignment [10,11]. In this work, we focus on the first aspect: *how to estimate human value preferences in a participatory system?*

We consider value preferences as a ranking of a given set of values. We estimate value preferences of individuals in a participatory system by incorporating both the *choices* made and *motivations* provided for those choices by the individuals. Estimating one's value preferences from both one's choices in a given context and the motivations given to support these choices might provide additional insights that could not be achieved considering only one source of information. But what should be done if these sources conflict? In other words, when estimating one's value preferences, what is more important: what one does (or chooses) or what one says?

We follow Samuel Scheffler's philosophical account that "valuing is deliberatively consequential" [12]. That is, if a user's choice is based on a deliberation of value preferences, the value preferences can be observed in the motivation the user provides for the choice [13,14,15]. Thus, valuing something involves a willingness to let it inform practical reasoning. For example, consider the value of online privacy. If a user, Alice, values privacy, she will deem reasons related to online privacy as important in related discussions, and will consider it a reason for action. Consider the context of indiscriminate use of social media (e.g., sharing potentially sensitive pictures with a large group of users). Alice may not encounter indiscriminate usage of social media often in her discussions. However, when she does, e.g., when discussing the topic with friends, we would expect her to explicitly mention privacy. On the other hand, although Alice considers privacy as important, she may prioritize social recognition over privacy when sharing a photo from her conference presentation. However, if Alice never considers online privacy as having deliberative relevance for action, it is unlikely that she values online privacy.

We propose three methods to address choice-motivation conflicts. These methods follow Scheffler's idea [12] by prioritizing value preferences observed in the motivations over value preferences estimated from the choices alone. We employ the proposed methods to estimate the value preferences of the participants of a large-scale survey on energy transition [16]. We evaluate the extent to which our methods' estimation concur with those of human evaluators. Our results show that addressing the inconsistencies between choices and motivations improves the estimation of value preferences.

We envision our work as supporting a *hybrid participatory system*, where humans participate in the decision-making process by making choices and providing motivations, and an AI agent supports the decision-making process by estimating the user's value preferences, as shown in Figure 1. The estimated value preferences can benefit (1) the policy maker by indicating both what users prefer and why, and (2) the user by unveiling the inconsistencies between their choices and motivation, thus helping them to clarify ambiguity and better articulate their (value) preferences.

Organization Section 2 discusses related works on estimating value preferences. Section 3 presents the context in which we study value preferences. Section 4 describes the methods we propose for value estimation, while Section 5 describes our experimental setting and Section 6 the results. Finally, Section 7 concludes the paper.

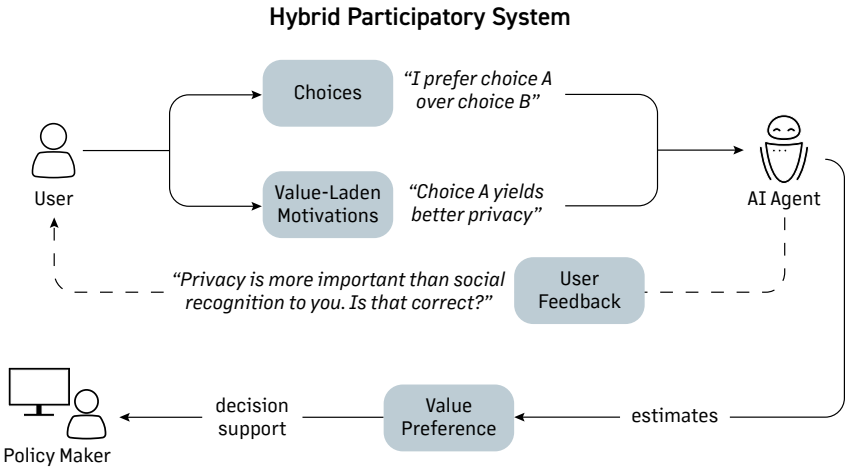


Figure 1. A hybrid participatory system, where human participants make choices and motivate those choices, and AI agents estimate the participants' value preferences to assist in decision making

2. Related works

We briefly discuss related works on estimating value preferences. Our work differs from these related works by estimating preferences from both choices and motivations.

2.1. Value preferences from surveys

Survey instruments such as the Portrait Value Questionnaire [6], Schwartz Value Survey [6], Value Living Questionnaire [17], and Moral Foundations Questionnaire [18] can be used to estimate one's preference towards a set of values. Further, some approaches combine self-reported surveys with participatory design [19,20], following the principles of Value Sensitive Design [21]. However, value questionnaires have been criticized for being incomplete and not context-sensitive [22]. In this work, we do not query participants directly about their values, but evaluate their choices and related motivation in context.

2.2. Value preferences from choices and behavior

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 [23]. For complex and high-dimensional environments, inverse reinforcement learning algorithms, which focus on extracting a "reward function" given observed optimal behavior, show promising results [24]. However, critiques on the infeasibility of estimating an individual's rationality and preferences (including value preferences) simultaneously [25] suggest the need for additional normative assumptions. We seek to address this critiques by incorporating textual motivations provided by humans for their choices.

2.3. Value preferences from text

A classical approach to value estimation from text is through value dictionaries—lists of word characteristic of certain values—by measuring the relative frequency of the words

describing each value [26] e.g., Moral Foundation Dictionary [27]. These dictionaries have been expanded through Natural Language Processing techniques [28,29,30], and limitations related to word count techniques have been approached via word embedding models [31,32]. Instead, other approaches use supervised machine learning on datasets annotated with value taxonomies [33,34,35,36]. In our work, we currently rely on human annotators to identify values on text, keeping the annotation process open for emerging and context-specific values. Although we envision potential integration with the aforementioned approaches, our work differs from them by focusing not only on the textual motivation provided to a set of choices, but also on the choices themselves.

3. Participatory Value Evaluation (PVE)

We estimate individual value preferences from choices and motivations provided via Participatory Value Evaluation (PVE) [3], an online participatory system. Specifically, we consider data from a PVE conducted between April and May 2020 involving 1,376 participants [16], aimed at supporting the municipality of Súdwest-Fryslân in the Netherlands in co-creating an energy policy, increasing citizen participation, and avoiding public resistance as happened in previous projects related to sustainable energy [37]. The main question to the citizens was: "What do you find important for future decisions on energy policy?" Six choice options (Table 1) were developed in consultation with 45 citizens. These options were presented in the PVE platform and the participants were asked to divide 100 points among the options. The choice options o_1 and o_2 were preferred more than other options. However, in most cases, participants distributed points to more than one option.

Table 1. Policy options in the Súdwest-Fryslân PVE

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

After dividing the points, the participants had the chance to motivate each of their choices. The values embedded in these textual motivations were later annotated by a set of four annotators using a grounded theory approach [38]. The annotators were first introduced to foundational concepts [6,27] and examples on 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. For simplicity, here we consider only a subset of this list, i.e. values mentioned at least 250 times across all project options in this work. Table 2 shows the value list we consider in our experiments.

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 ($\sim$76%) times for o_3, and only 3 times for o_6.

Table 2. Selected values for the Súdwest-Fryslân PVE

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 it is 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 3. Distribution of values annotated for options

		Options						
		o_1	o_2	o_3	o_4	o_5	o_6	O
Annotated values	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**	

4. Methods

Our goal is to estimate an individual's value system from the division of points across a set of policy options (*choices*) and the textual *motivations* provided to each choice. As the choices and motivations were provided within a specific context (energy transition), the resulting value system is intended to represent the individual within such context.

4.1. Value system

Values can be ordered according to their subjective importance as guiding principles [6]. 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 preference via a ranking [39]. Adapting from [40], 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, an R is the ranking of V which represent 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 $v_a \succ v_b$.*

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. In this work,

we focus on total orders as an initial step on estimating value preferences, given existing challenges on fairly aggregating partial orders [41].

4.2. 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 to these choices. Let $O = \{o_1, \ldots, o_n\}$ be a set of options i can choose from in a specific context (for example, the policy options presented in Table 1). We assume that i indicates her preferences, C^i, among the choices in O by distributing a certain number of points, p, among the options in O.

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

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

$$M^i = \{m_1, \ldots m_n\}, \text{ where } m_i = \varnothing \text{ if } c_i = 0$$

Further, following the premise that valuing is deliberatively consequential, if an individual's value system influences her choice c_i, we expect her to mention the values which support choice c_i in the motivation provided. Thus, here we represent motivations m_i as the set of values that are mentioned in them (for example, the values in Table 2):

$$m_i = \{v_1, \ldots v_m\}, \text{ if } v_i \text{ influenced } c_i, \quad v_i \in V$$

4.3. Value-option matrix

Consider that the values relevant in choosing each option o_i can be determined a priori.

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

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

VO is the starting point for computing individual value systems, as it represents an initial guess of value preferences in the energy transition context based on the available choices by all participants. Thus, we initialize each individual's *VO* matrix (VO^i) as:

$$VO^i = VO \tag{1}$$

Figure 2 shows the relation between choices, motivations, values, and *VO* matrix.

4.4. Estimating value preferences

Given VO^i, we propose multiple methods to estimate an individual's value system $\langle V, R^i \rangle$ from (1) choices (method *C*), (2) motivations (method *M*), or (3) choices and motivations (methods *TB*, *MC*, and *MO*). We provide the rationale behind each method in the related subsection. The methods *C* and *M*, which use either choices or motivation, are used as baseline for evaluating the other three methods (*TB*, *MC*, and *MO*). All methods can be applied sequentially—however, the order in which they are applied can change the final ranking. Figure 3 shows the main elements of each method, which are described next.

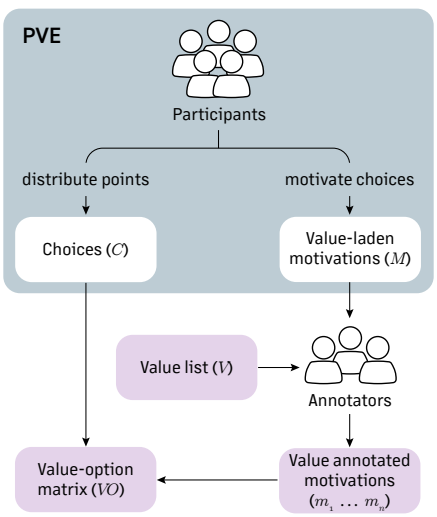


Figure 2. Relationship between choices, motivations, and values

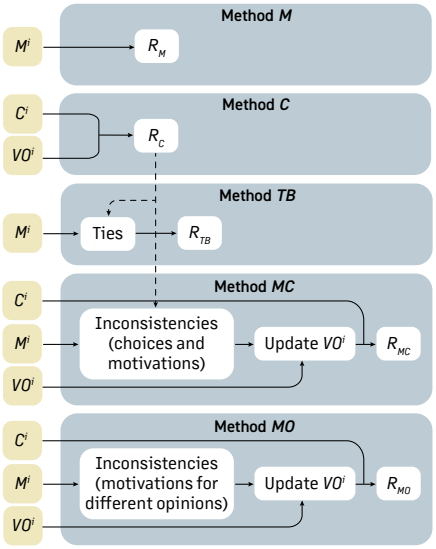


Figure 3. Overview of the proposed methods

4.4.1. Method C

To estimate an individual's value ranking R^i_C solely based on her choices C^i (vector of size $|O|$, i.e., number of options), we assume that the individual's choices completely align with her value system. First, we compute the importance of values (U^i) for the individual by weighing the values supported by each option with the points (c_i) 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.

$$U^i = VO^i \times C^{i^T} \tag{2}$$

$$R^i_C = \texttt{rank}(U^i) \tag{3}$$

4.4.2. Method M

To estimate an individual's value ranking R^i_M solely based on the motivations M^i provided to her choices C^i, first we 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.

4.4.3. Method TB: Motivations as tie breakers

We use the motivations M^i as *tie breakers* with the goal of reducing indifferent preferences in a value system. 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 ($v_a \sim v_b$). Due to symmetry, we consider that $\tau_{a,b} = \tau_{b,a}$.

If there is a tie $\tau_{a,b}$ and if 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 ($\tau_{a,b} \in R^i$)
  if $\left(\exists m \in M^i : v_a \in m\right) \wedge \left(\nexists m \in M^i : v_b \in m\right)$ **then**
    set $v_a \succ v_b$ in R^i_{TB} ;
  else if $\left(\exists m \in M^i : v_b \in m\right) \wedge \left(\nexists m \in M^i : v_a \in m\right)$ **then**
    set $v_b \succ v_a$ in R^i_{TB} ;

4.4.4. 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 her motivations. That is, R^i indicates $v_b \succ v_a$ but v_a is supported in a motivation $m_o \in M^i$ 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.

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 supported by $m_o = \{v_a\}$ to 0. 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}
for ($m_o \in M^i$)
  for ($v_a \in m_o$)
    for ($v_b \in V \setminus \{v_a\}$)
      if $v_a \prec v_b$ **then**
        $VO^i(v_b, o) = 0$;
$U^i = VO^i \times C^i$;
$R^i_{MC} = \texttt{rank}(U^i)$;

4.4.5. Method MO: Motivations are only relevant for one option

The motivations M^i provided for different options can also bring inconsistencies. For example, assume options o_1 and o_2, for which all values $v_i \in V$ are considered relevant. 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 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 ($m_a \in M^i : m_a \neq \varnothing$)
  for ($m_b \in M^i \setminus \{m_a\}$)
    $V_\alpha = V \setminus \{v : v \in m_a\} : VO^i(v, o_a) == 1$; /* Values supporting o_a in VO^i, except values in m_a */
    for ($v_x \in V_\alpha$)
      if $v_x \in m_b$ **then**
        for ($v_y \in m_a$)
          $V_\beta = V \setminus \{v : v \in m_b\} : VO^i(v, o_b) == 1$; /* Values supporting o_b in VO^i, except values in m_b */
          if $v_y \in V_\beta$ **then**
            $VO^i_{MO}(v_x, o_a) = 0$;
$VO^i \leftarrow VO^i_{MO}$;
$U^i = VO^i \times C^i$;
$R^i_{MO} = \texttt{rank}(U^i)$;

5. Experimental setting

Given the participation process on energy transition using PVE [16] described in Section 3, we consider a given value v as relevant for option o if at least t motivations (in our case, we considered $t = 20$) among all participants were annotated with v for o. *VO* (as described in section 4) is then represented by Table 4.

Table 4. Value-option matrix (*VO*) for the energy transition PVE

Values \ Options	o_1	o_2	o_3	o_4	o_5	o_6
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	1	0	1
v_5	1	1	0	0	1	0

5.1. Methods for evaluation

We analyze each method (*C*, *M*, *TB*, *MC*, and *MO*) individually, and a sequential combination of all 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 start with *MC* because it addresses inconsistencies within the same participant (which happens more frequently), and then continue with *MO* (less frequent). To combine these methods sequentially we use the ranking resulting from *MO* as input for *MC*, and the ranking resulting from *MC* as input for *TB*. Finally, for the indi-

vidual 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 ranking.

5.2. Evaluation procedure

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

To simplify the task, the evaluators were sequentially presented with a participant's choices and motivations, proposed pairs of values (e.g., v_1 and v_2), and asked to compare the two values with the following options: (1) $v_1 \succ v_2$; (2) $v_1 \prec v_2$; (3) $v_1 \sim v_2$; or (4) "I do not know", if they believe there is not enough information to make a proper comparison. This comparison was repeated up to four times per each selected participant, 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 could estimate value preferences similarly to the human evaluators. Within the application context illustrated in Figure 1, we expect that as the methods' rankings mirror human intuition, they might provide meaningful feedback to participants in a participatory system.

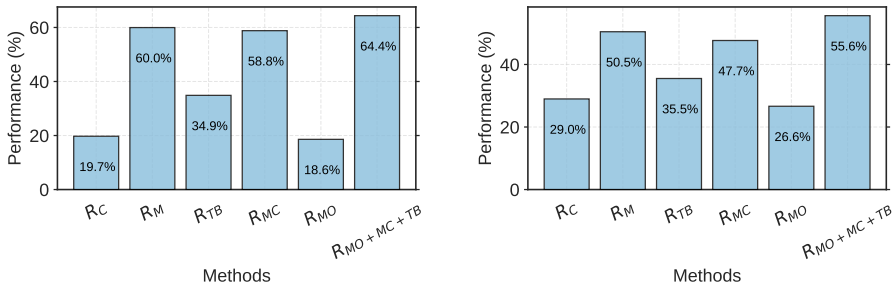
6. Results and discussion

In this section we present and discuss the results of our methods. We aim to answer two questions: (1) How well can each method estimate value systems compared to humans? (2) How does the estimation of value systems differ among the methods proposed?

6.1. Evaluation

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. Figure 4 presents 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%) of the comparisons, resulting in a Kappa score of 0.247, which is considered a fair agreement [42]. To mitigate the effect of individual biases, in the remainder of the analysis we focus on the pairwise comparisons where evaluators agreed on, as presented in Figure 5.

As both Figures 4 and 5 display, the rankings R_M, R_{MC} and $R_{MO+MC+TB}$ provide the best performance in terms of human-like value estimation. When compared to R_C, the combined method $R_{MO+MC+TB}$ estimated value system 3.31 times more similarly to humans (considering the subset where evaluators agreed). Further, we observe that R_M and R_{MC} also performs better than R_C. The only exception in terms of performance is



(a) Evaluator 1 (b) Evaluator 2

Figure 4. Performance of the methods, measured as the overlap with each evaluator's answers

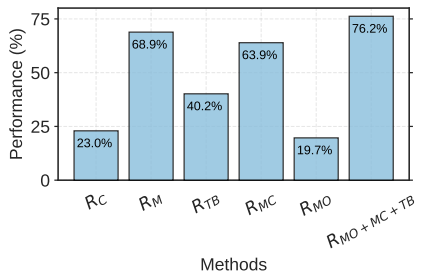


Figure 5. Performance of the methods, measured as the overlap with the answers where evaluators agree

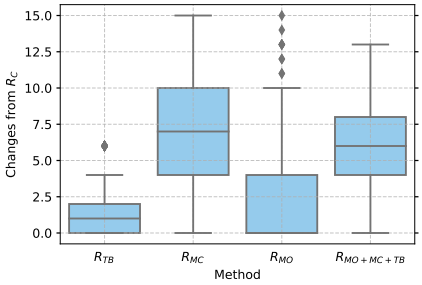


Figure 6. Average changes in the rankings when compared to R_c

R_{MO}, which performs slightly worse than R_C. These findings show that the combined use of choices and motivations in estimating value preferences can significantly increase the degree to which an automated method can estimate a value system similarly to humans.

6.2. 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: $v_1 \succ v_2 \succ v_4 \succ v_5 \succ v_3$
- R_M: $v_3 \succeq v_1 \succeq v_2 \succeq v_5 \succ v_4$
- R_{TB}: $v_1 \succ v_2 \succ v_4 \succ v_5 \succ v_3$
- R_{MC}: $v_1 \succ v_2 \succeq v_3 \succeq v_4 \succeq v_5$
- R_{MO}: $v_1 \succ v_2 \succ v_4 \succ v_5 \succ v_3$
- $R_{MO+MC+TB}$: $v_1 \succ v_2 \succ v_3 \succeq v_4 \succeq v_5$

Method C ranked the value v_1 as the most important for all individuals, regardless of their choices, due to the characteristics of the value option-matrix (VO) 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_C^i 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).

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

7. Conclusion and future works

We introduce methods for an AI agent to estimate value preferences of individuals in a hybrid participatory system from one's choices and value-laden motivations, with the goal of generating a partially 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 systems improves the performance of value estimation by more than three times (in terms of similarity to human evaluators' value estimation) and yields preferences that are more individually-tailored.

In future experiments, participants themselves could provide direct feedback to the AI agent, instead of relying on external evaluators. Further, Natural Language Processing algorithms, e.g., [32,33,35], could be used to scale-up experiments by automatically identifying the values supporting the motivations. Finally, we suggest exploring other approaches to associate values with choice options beyond a binary matrix, since values can have different ethical impacts in different contexts.

Our methods can be used to support policy-makers by aggregating individual value systems into a societal value system [43] to provide an overview on the value preferences of a population. Further, as our work is focused on estimating value systems from choices and motivations, it has the potential to contribute to value alignment between AI and humans in a hybrid participatory system. By connecting an individual's motivations to her values in a transparent manner, we support the development of systems that empower humans in asserting their values and contribute to meaningful human control over AI.

Acknowledgments: This work was partially supported by TU Delft's AiTech initiative and by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] Dallhammer E, Gaugitsch R, Neugebauer W, Böhme K. Spatial Planning and governance within EU policies and legislation and their relevance to the New Urban Agenda. European Committee of the Regions: Bruxelles, Belgium. 2018;.

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

[3] Mouter N, Hernandez JI, Itten AV. Public participation in crisis policymaking. How 30, 000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures. PLoS ONE. 2021;16(5):1–42.

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

[5] Iandoli L, Quinto I, De Liddo A, Buckingham Shum S. 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. 2016;67(5):1052–1067.

[6] Schwartz SH. An Overview of the Schwartz Theory of Basic Values. Online Readings in Psychology and Culture. 2012;2(1):1–20.

[7] Dignum V. Responsible Autonomy. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17); 2017. p. 4698–4704.

[8] Liscio E, van der Meer M, Siebert LC, Jonker CM, Mouter N, Murukannaiah PK. Axies: Identifying and Evaluating Context-Specific Values. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. AAMAS '21. Online: IFAAMAS; 2021. p. 799–808.

[9] Liscio E, van der Meer M, Siebert LC, Jonker CM, Murukannaiah PK. What Values Should an Agent Align With? Autonomous Agents and Multi-Agent Systems. 2022;36(23):32.

[10] Dignum V. Responsible Artificial Intelligence: Designing AI for Human Values. ITU Journal: ICT Discoveries. 2017;(1):1–8.

[11] Murukannaiah PK, Ajmeri N, Jonker CJM, Singh MP. New Foundations of Ethical Multiagent Systems. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. AAMAS '20. Auckland: IFAAMAS; 2020. p. 1706–1710.

[12] Scheffler S. Valuing. In: Equality and Tradition: Questions of Value in Moral and Political Theory. 1st ed. Oxford University Press; 2012. p. 352.

[13] Dietz T, Stern PC. Toward a theory of choice: Socially embedded preference construction. Journal of Socio-Economics. 1995;24(2):261–279.

[14] Kenter JO, Reed MS, Fazey I. The deliberative value formation model. Ecosystem Services. 2016;21:194–207.

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

[16] Spruit SL, Mouter N. 1376 residents of Súdwest-Fryslân about the future energy policy of their municipality: the results of a consultation; 2020. Available from: https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan/.

[17] Wilson KG, Sandoz EK, Kitchens J, Roberts M. The valued living questionnaire: Defining and measuring valued action within a behavioral framework. Psychological Record. 2010;60(2):249–272.

[18] Graham J, Nosek BA, Haidt J, Iyer R, Koleva S, Ditto PH. Mapping the moral domain. Journal of personality and social psychology. 2011;101(2):366.

[19] Liao QV, Muller M. Enabling Value Sensitive AI Systems through Participatory Design Fictions. arXiv; 2019.

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

[21] Friedman B, Hendry DG. Value Sensitive Design: Shaping Technology with Moral Imagination. The MIT Press; 2019.

[22] Boyd RL, Wilson SR, Pennebaker JW, Kosinski M, Stillwell DJ, Mihalcea R. Values in words: Using language to evaluate and understand personal values. In: Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015; 2015. p. 31–40.

[23] Benabou R, Falk A, Henkel L, Tirole J. Eliciting Moral Preferences: Theory and Experiment. Princeton University; 2020.

[24] Russell SJ. Human Compatible: AI and the Problem of Control. 1st ed. Viking Press; 2019.

[25] Mindermann S, Armstrong S. Occam's razor is insufficient to infer the preferences of irrational agents. In: NeurIPS. 2676; 2017. p. 1–12.
[26] Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC. Mahway: Lawrence Erlbaum Associates. 2001;.
[27] Graham J, Haidt J, Koleva S, Motyl M, Iyer R, Wojcik SP, et al. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In: Advances in Experimental Social Psychology. vol. 47. Amsterdam, the Netherlands: Elsevier; 2013. p. 55–130.
[28] Rezapour R, Shah SH, Diesner J. Enhancing the measurement of social effects by capturing morality. In: Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2019. p. 35–45.
[29] Hopp FR, Fisher JT, Cornell D, Huskey R, Weber R. The Extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text. Behavior Research Methods. 2020;p. 1–23.
[30] Araque O, Gatti L, Kalimeri K. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. Knowledge-Based Systems. 2020;191(3):105184.
[31] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013;26:3111–3119.
[32] Garten J, Hoover J, Johnson KM, Boghrati R, Iskiwitch C, Dehghani M. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior research methods. 2018;50(1):344–361.
[33] Mooijman M, Hoover J, Lin Y, Ji H, Dehghani M. Moralization in social networks and the emergence of violence during protests. Nature human behaviour. 2018;2(6):389–396.
[34] Hoover J, Portillo-Wightman G, Yeh L, Havaldar S, Davani AM, Lin Y, et al. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science. 2020;11(8):1057–1071.
[35] Liscio E, Dondera AE, Geadau A, Jonker CM, Murukannaiah PK. Cross-Domain Classification of Moral Values. In: Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '22. Seattle, USA: Association for Computational Linguistics; 2022. p. 1–13. To appear.
[36] Kiesel J, Alshomary M, Handke N, Cai X, Wachsmuth H, Stein B. Identifying the Human Values behind Arguments. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. ACL '22. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 1–13.
[37] Dutch Ministry of Economic Affairs and Climate. National Climate Agreement-The Netherlands; 2019.
[38] Heath H, Cowley S. Developing a grounded theory approach: a comparison of Glaser and Strauss. International Journal of Nursing Studies. 2004 2;41(2):141–150.
[39] Zintgraf LM, Roijers DM, Linders S, Jonker CM, Nowé A. 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; 2018. p. 1477–1485.
[40] Serramia M, Rodríguez-Aguilar JA, Lopez-Sanchez M. A Qualitative Approach to Composing Value-Aligned Norm Systems. In: Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems. Online: IFAAMAS; 2020. p. 9–13.
[41] Pini MS, Rossi F, Venable KB, Walsh T. Aggregating partially ordered preferences: impossibility and possibility results. In: Proceedings of the 10th conference on Theoretical aspects of rationality and knowledge; 2005. p. 193–206.
[42] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159.
[43] Lera-Leri R, Bistaffa F, Serramia M, Rodríguez-Aguilar J. Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p-Regression. In: Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems. Auckland, New Zealand: IFAAMAS; 2022. p. 1–9.