Axies: Identifying and Evaluating Context-Specific Values

Enrico Liscio
Delft University of Technology
E.Liscio@tudelft.nl

Michiel van der Meer
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Luciano C. Siebert
Delft University of Technology
L.CavalcanteSiebert@tudelft.nl

Catholijn M. Jonker
Delft University of Technology
C.M.Jonker@tudelft.nl

Niek Mouter
Delft University of Technology
N.Mouter@tudelft.nl

Pradeep K. Murukannaiah
Delft University of Technology
P.K.Murukannaiah@tudelft.nl

ABSTRACT

The pursuit of values drives human behavior and promotes cooperation. Existing research is focused on general (e.g., Schwartz) values that transcend contexts. However, context-specific values are necessary to (1) understand human decisions, and (2) engineer intelligent agents that can elicit human values and take value-aligned actions.

We propose Axies, a hybrid (human and AI) methodology to identify context-specific values. Axies simplifies the abstract task of value identification as a guided value annotation process involving human annotators. Axies exploits the growing availability of value-laden text corpora and Natural Language Processing to assist the annotators in systematically identifying context-specific values.

We evaluate Axies in a user study involving 60 subjects. In our study, six annotators generate value lists for two timely and important contexts: Covid-19 measures, and sustainable Energy. Then, two policy experts and 52 crowd workers evaluate Axies value lists. We find that Axies yields values that are context-specific, consistent across different annotators, and comprehensible to end users.

KEYWORDS

Values; Ethics; Context; Natural Language Processing

ACM Reference Format:
Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 10 pages.

1 INTRODUCTION

Values are abstract motivations that justify opinions and actions, and are intrinsically linked to feelings and goals [35]. As agents operate in sociotechnical systems [27] on behalf of and among humans [2], agents' behavior must accord with human values.

There is a growing recognition [33, 38] that values are central to robust and beneficial AI. In a value-sensitive AI system, an agent must first elicit or learn the value preferences of the stakeholders [4, 37]. Then, the agent can reason about aligning its actions with the values of the stakeholders [1, 8, 9, 23]. However, a crucial question that must be answered before these steps is:

> **What values** should an agent elicit, learn, or align with?

Several lists of *general values* have been proposed by ethicists [31, 35], political scientists [13], designers [10] and, recently, computer scientists [46]. These value lists aim to be applicable, broadly, across cultures and contexts. However, for concrete analysis and use of values, e.g., to (1) elicit stakeholders' values [18], (2) communicate values to stakeholders [45], (3) translate values into design requirements [29, 43], (4) reason about conflicting values [1, 28], (5) align values and norms [36], and (6) verify value adherence of an AI system [42], values must be situated within a context.

We define a *context-specific value* as a value that is applicable and defined specifically within a context. For example, in the context of information sharing on Social Media, privacy is an applicable value, but physical health is likely not (unless we are talking about the health effects of Computer Use, which is another context). Further, privacy can be interpreted as intruding one's solitude, or control on information collection, processing, and dissemination [39]. Thus, privacy defined as one's ability to control the extent to which her information is collected, processed, and disseminated is a value specific to the context of social Media.

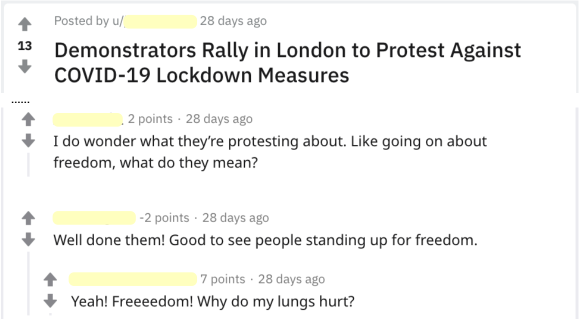
How can we identify values specific to a context? Since values are (high-level) cognitive abstractions, human intelligence is necessary to conceptualize a value and reason about its relevance to a context. However, thinking about values is challenging even for humans [18, 29]. Thus, we need to systematically guide and assist humans in the process of identifying context-specific values.

We propose Axies (from the Greek word *αξίες*, meaning *values*), a hybrid (human and AI) methodology to engage humans in identifying context-specific values and support the process via Natural Language Processing (NLP) techniques. A key idea behind Axies is to simplify the abstract task of value identification to a concrete task of value annotation given a (textual) value-laden opinion. With this approach, Axies enables human annotators to (1) learn about a context by exploring opinions about the context, and (2) think about values one opinion at a time.

There is a growing availability of value-laden opinions for many contexts on the Web, e.g., on discussion forums, tweets, and blogs. For example, Figure 1 shows examples of value-laden opinions on a Reddit discussion forum. By showing this opinion, Axies triggers a value annotator to think about the values of freedom and health in the context of Covid-19 measures. Value-laden opinions can also be collected by explicitly consulting a target population, e.g., [25].

Annotating a large opinion corpus is a significant effort. First, Axies distributes this task among a small group of annotators. Inspired by traditional coding methods such as the grounded theory method [12], the annotators engage in both divergent and convergent thinking by individually exploring the opinion corpus and

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.



Posted by u/[redacted] 28 days ago

13

Demonstrators Rally in London to Protest Against COVID-19 Lockdown Measures

......

[redacted] 2 points · 28 days ago

I do wonder what they're protesting about. Like going on about freedom, what do they mean?

[redacted] -2 points · 28 days ago

Well done them! Good to see people standing up for freedom.

[redacted] 7 points · 28 days ago

Yeah! Freeeedom! Why do my lungs hurt?

...... Source: https://www.reddit.com/r/CoronavirusUK/comments/iisk44/demonstrators_rally_in_london_to_protest_against/

Figure 1: Example value-laden opinions on a Reddit forum

collaboratively consolidating a value list. Second, Axies employs an active learning strategy [5] to control the order in which opinions are shown to the annotators to reduce the annotation effort.

We conduct a user study of 60 subjects to answer three questions.

Specificity Does Axies yield *context-specific* values?
Comprehensibility Does Axies yield *clear* and *distinct* values?
Consistency Does Axies yield a *consistent* set of values, independent of the people applying the methodology?

In our study, first, six annotators (in two groups of three) generate value lists specific to two contexts: Covid-19 relaxation measures, and sustainable Energy policies. Second, two policy experts evaluate the context-specificity of Covid and Energy value lists. Finally, we employ 52 crowd workers to evaluate the comprehensibility of Axies value lists, and the consistency between value lists generated by different annotator groups for the same context.

Contributions (1) We propose Axies, a hybrid methodology to guide a group of human annotators in identifying context-specific values. Axies employs NLP techniques and active learning to engage the annotators in inducing values from an opinion corpus. (2) We conduct three experiments, generating four value lists for two contexts, and demonstrate that Axies yields context-specific, comprehensible, and consistent value lists.

Organization Section 2 reviews related works. Section 3 describes Axies. Section 4 describes the three experiments. Section 5 discusses our results. Section 6 concludes the paper. We include code, study protocols, and data as supplemental material [20, 21].

2 RELATED WORKS

Values can be expressed through language, behavior, and customs. Values vary significantly across people, socio-cultural environments, and contexts [9]. Thus, ascertaining values requires extensive personal communication and analysis. However, the burst of online communication and social media provides an unprecedented opportunity for scientists to study several social phenomena [3], including value understanding and estimation from language.

Values in Words. NLP techniques allow the (semi-) automatic estimation of values from text. Liu et al. [22] present a psychographic analysis of values based on users' word use from e-commerce reviews. However, since moral values are often only implicit in language, automated extraction of values from text is challenging. Lin et al. [19] estimate moral values in tweets by combining textual features and background knowledge (context) from Wikipedia. Hoover et al. [16] use a Distributed Dictionary Representation [11] to study the expression of moral values in tweets about charitable donations posted during and after Hurricane Sandy.

The works above start from a value list: [22] uses values from Schwartz Value Survey [35] and [16, 19] use Moral Foundations Dictionary [13]. In contrast, our objective is to identify a value list.

Identifying Values. Boyd et al. [6] demonstrate that values learned from free-response language (e.g., Facebook status messages) yield better predictive coverage of real-world behavior than values extracted from self-report questionnaires such as Schwartz Value Survey. Building on [6], Wilson et al. [46] describe a crowd-powered algorithm to generate a hierarchy of general values. Teernstra et al. [41] demonstrate that a text classifier (of Twitter discussions) predicts values from Moral Foundations Theory more accurately than a hand-crafted dictionary of general value-related keywords.

Similar to the works above, we employ a data-driven approach toward values. Unlike these approaches (which consider general values), we focus on context-specific values essential for concrete use and analysis of values, e.g., [1, 18, 28, 29, 36, 42, 43, 45].

Value Sensitive Design (VSD). Value identification is central to VSD [10], a broad set of methods for designing technology that accounts for human values. VSD includes methods for identifying value sources, representing values, and resolving value tensions. Pommeranz et al. [29] argue that an essential step in effective realization of VSD is the instantiation of abstract values in specific contexts. They also recognize the need for self-reflection triggers since reflecting on values is not natural to most people.

Axies fills the gaps in VSD Pommeranz et al. [29] recognize. First, Axies targets the identification of context-specific values. Second, Axies provides concrete triggers (opinions) to human annotators (who need not be design experts) for reflecting on values.

3 AXIES METHODOLOGY

Figure 2 shows an overview of the Axies methodology. Given a context-specific opinion corpus, Axies yields a context-specific value list applicable to the *users* producing the opinion corpus. To do so, Axies (1) exploits NLP techniques; and (2) engages a group of value *annotators* in the systematic steps of exploration (individual) and consolidation (collaborative).

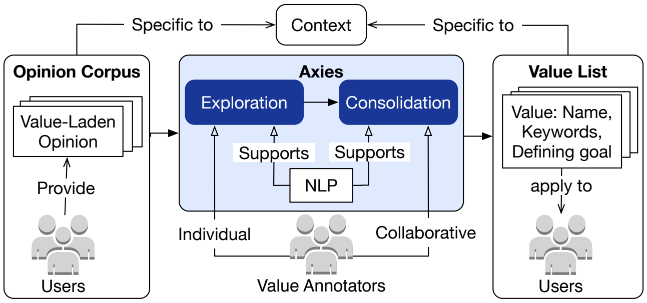


Figure 2: Overview of the Axies methodology

3.1 Opinion Corpus

The input to Axies is a corpus of users' opinions within a context. Axies requires the corpus to include *value-laden* opinions. A value-laden opinion indicates a user's value, explicitly or implicitly (e.g., in Figure 1 freedom is explicit and health is implicit).

3.1.1 *Participatory Value Evaluation (PVE).*

We construct the opinion corpora for Axies evaluation (Section 4) using data from PVE. A PVE elicits citizens' preferences about government policy options [24]. Specifically, participants are offered a predetermined set of policy options and informed about impacts. Then, participants are to advise their preferred portfolio of options while respecting the constraints of the government and (optionally) provide motivations for their choices.

A PVE participant's motivation is included as an opinion in our corpus. Often, these opinions offer valuable insights into the values of PVE participants. Table 1 shows examples of value-laden opinions of participants in a recent PVE on Covid-19 relaxation measures in the Netherlands [25].

Table 1: Example value-laden opinions in a Covid-19 PVE

Preference	Motivation
Nursing homes allow visitors again	Loneliness and isolation are a bigger killer than Corona.
All restrictions are lifted for persons who are immune	Someone's got to keep the economy going.

3.2 Value List

The output of Axies is a *value list* specific to the context in which an opinion corpus is produced and applicable to the users producing the corpus. We represent each value in the list by a name, a set of *keywords* that characterize the value in the context, and a *defining goal* [35] that specifies what "holding a value" means in that context. For instance, Table 2 shows example Covid-19 specific values, applicable to Dutch citizens, produced in the Axies evaluation.

Table 2: Examples of Dutch citizens' Covid-19 values

Name	Keywords	Defining goal
Mental health	Loneliness, quality of life, stress	The strive towards protecting and improving one's emotional and psychological well-being.
Economic prosperity	Economy, stability, bankruptcy	Being able to pay and afford what you need.

3.3 Value Annotators

Axies is intended to be executed by a small group of annotators, who (1) produce individual value lists during *exploration*, and (2) collaboratively merge the individual lists during *consolidation.*

Axies facilitates *inductive reasoning* in that the annotators infer values held by users (theory) based on the opinions users express (evidence). A key advantage of this inductive approach is that Axies yields values grounded in data. In addition, the inductive process provides an opportunity to systematically guide the annotators.

3.4 Axies: Value Exploration

In the exploration phase, each annotator independently develops a value list (with name and keywords for each value) by analyzing users' opinions. Depending on the context, opinion corpora can be quite large. For example, the Covid-19 opinion corpus [25] we evaluate contains about 60,000 opinions. Thus, it is not feasible for an annotator to analyze each opinion in a corpus.

Axies seeks to (1) reduce the number of opinions each annotator analyzes to produce a stable value list, and (2) increase the coverage of opinions (with respect to the corpus) the group of annotators analyze. To achieve these objectives, Axies employs NLP and active learning techniques to control the order in which the opinions in the corpus are exposed to the annotators. Thus, each annotator analyzes only a subset of the opinions in the corpus.

3.4.1 *Opinion and Value Embeddings.*

Axies represents opinions and values as vectors computed from the Sentence-BERT [30] sentence embedding model M. M takes as input a word or a sentence and returns its vector representation in an n-dimensional space ($n = 768$, in our case). Let $M(o)$ be the vector representation of an opinion o. Let n_v be the name and $K_v = \{k_v^1, \ldots, k_v^n\}$ be the set of keywords of a value v. Then, Axies computes the value vector $M(v)$ using the Distributed Dictionary Representation [11] as:

$$M(v) = \frac{M(n_v) + \sum_{k \in K_v} M(k)}{||M(n_v) + \sum_{k \in K_v} M(k)||}. \tag{1}$$

With the vector representations, we can compute cosine similarity between values and opinions during opinion selection.

3.4.2 *Procedure.*

Let A be a set of value annotators for a context. Then, each annotator $a \in A$ follows the exploration steps below.

Opinion selection Axies employs an active learning technique known as *Farthest First Traversal* (FFT) [5, 32]. Using FFT, Axies selects opinions such that an opinion shown to an annotator a is the farthest from the opinions already shown to the annotators in group A and the values already annotated by the annotator a.

Algorithm 1 shows the pseudocode for selecting an opinion to show an annotator a. We run one instance of this algorithm to select opinions for all annotators in A to reduce the overlap in opinions shown to different annotators in A (thereby, increasing the coverage of opinions shown to the annotators in A). However, for each annotator $a \in A$, we employ the individual value list, V_a.

Annotation Algorithm 1 shows opinions to an annotator, sequentially. After seeing an opinion, an annotator can add a value (with a name and keywords), or update the name or keywords of an existing value in their value list. The annotators are asked to reason about the values underlying a user's opinion. However, the value name or keywords need not explicitly appear in the opinion.

When an annotator adds a value name, we show as keyword suggestions to the annotator the five most similar words to the value name based on a counter-fitted word embedding model [26], trained to push synonyms closer and antonyms farther.

Algorithm 1: Fetching next opinion using FFT

```
**Input:** $O, M$ ;        /* Opinions, Embedding model */
**Output:** $V_a$ ;                     /* Value list of $a$ */
initialization: $\forall o \in O : d_o = \infty; V_a = \emptyset$;
**while** $O \neq \emptyset$ && $\neg$saturated($V_a$) **do**
    $o_{\text{next}} = \arg\max_{o \in O} d_o$ ;  /* break ties randomly */
    $O = O - o_{\text{next}}$;
    $V_a^{\text{old}} = V_a$;
    update_values($V_a, o_{\text{next}}$);
    $V_a^{\delta} = V_a - V_a^{\text{old}}$;
    $\forall o \in O : d_o =$
      $\min \begin{Bmatrix} d_o, \\ \texttt{cosine\_distance}(M(o), M(o_{\text{next}})), \\ \forall v \in V_a^{\delta} : \texttt{cosine\_distance}(M(o), M(v)) \end{Bmatrix}$;
**end**
```

Termination An annotator must judge when to stop annotating. We suggest the annotators to reach *inductive thematic saturation* [34], i.e., to continue annotation until the value list incurs no new changes for several new opinions shown to the annotator.

We show a *progress plot*, similar to Figure 3, to assist the annotators in deciding on termination. The progress plot shows a bar for each opinion seen by an annotator; the length of the bar is the FFT distance (d_o) at which the opinion was fetched; and the bar color indicates the annotator's action after seeing the opinion. A long sequence of opinions without addition of value names or keywords is an indicator of a stable value list.

Refinement Finally, Axies can fetch opinions similar to a value by computing cosine similarity between a value and the opinions not yet shown to an annotator. An annotator can fetch opinions similar to a value to refine the value, especially if it is not well formulated. Such a phase is visible in the final gray bars in Figure 3.

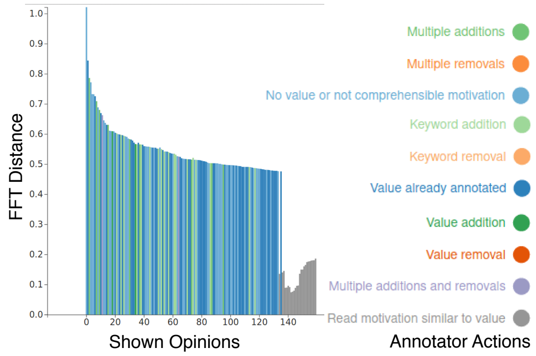


Figure 3: Example progress plot of exploration

3.5 Axies: Value Consolidation

In the consolidation phase, the annotators in group A collaborate and combine their individual value lists. Exploration and consolidation are complementary in that exploration facilitates divergent thinking whereas consolidation facilitates convergent thinking.

3.5.1 Procedure. To facilitate consolidation, Axies creates a combined value list, $V_A = \bigcup_{a \in A} V_a$ (the union of individual value lists), and guides the annotators in systematically refining V_A.

Value pairs To simplify the consolidation process, Axies requires the annotators to consolidate only a pair of values at a time. Yet, consolidation is cognitively challenging. If performed naively, the annotators must compare all possible pairs of values in V_A, and repeat that process several times, to arrive at a refined V_A.

To reduce the cognitive load, Axies controls the order in which value pairs are presented to the annotators—the most similar value pair from V_A is shown first. This approach is beneficial because similar values are likely to be merged, reducing the size of V_A, which in turn, reduces the number of value pairs to consolidate.

Consolidation actions Given a pair of values, the original annotator of each value in the pair describes the value to the other annotators in the group. Axies can fetch the opinions that led to the annotation of a value to assist an annotator in recalling the reasoning behind the annotation. The annotators in the group discuss whether the two values are conceptually the same or distinct. Accordingly, the annotators can take one of the following actions.

- **Merge** the two values, if they are conceptually identical. The annotators may choose one of the two names or a new name for the merged value, and retain or update the keywords.
- **Update** one or both values, if the values are conceptually distinct, but changes in name or keywords make the distinction clearer.
- **Take no action**, if the two values are conceptually distinct, and the distinction is clear as is. If the annotators take no action for a pair of values, that pair is not shown to the annotators again even if that is the most similar value pair in V_A.

Termination Terminating consolidation is subject to annotators' judgment as to whether the value list requires further refinement or not. Axies shows a plot (similar to Figure 4) for the annotators to keep track of progress. As shown in the plot, the pairs of similar values shown early in the consolidation process lead to several value updates and merges. However, annotators may also manually choose values to merge or update; the intermittent spikes in Figure 4 are due to such manual choices.

Reflection As the final step, the annotators critically reflect on the consolidated value list. In particular, Axies suggests the annotators to analyze each value in the list with respect to the main features of values. Schwartz [35, p3] describes six main features of values;

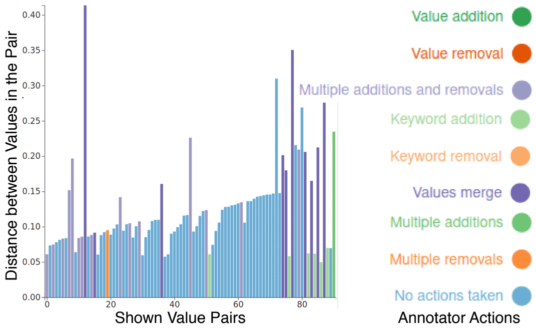


Figure 4: Example progress plot of consolidation

we include five of those, excluding the feature that (basic) values "transcend contexts" since Axies aims for context-specific values.

During reflection, Axies also asks the annotators to add a defining goal for each value in the list. The defining goal characterises what "holding a value" means. That is, a person holding a value in a context is likely to have the corresponding goal in that context. We defer the task of adding defining goals till the end of consolidation so that the task can be performed once for the final list of values.

4 EXPERIMENTS

We conducted three experiments, involving a total of 60 human subjects, to evaluate Axies as shown in Figure 5. These experiments were approved by our university's Ethics Committee, and we received an informed consent from each subject.

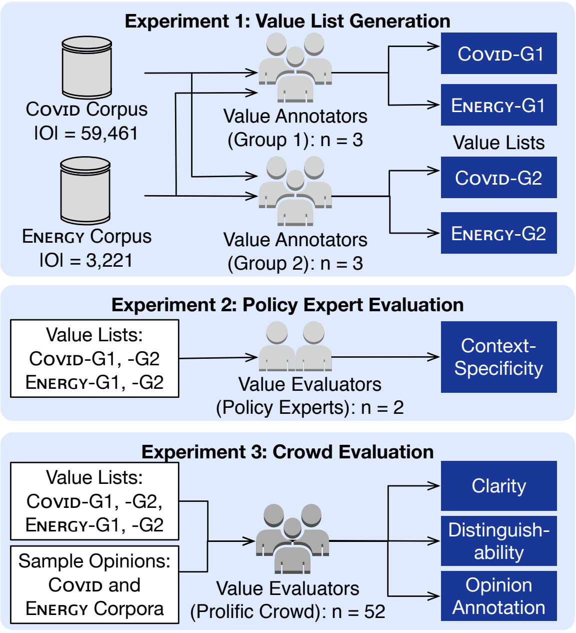


Figure 5: Overview of our experimental setup

In Experiment 1, two groups, G1 and G2, of three annotators each, employ Axies to generate value lists for two contexts (Covid, and Energy). Let these lists be Covid-G1, Energy-G1, Covid-G2, and Energy-G2. We employ these lists in the other two experiments to answer our three research questions on Axies:

Specificity In Experiment 2, we analyze the context specificity of Covid (G1 and G2), and Energy (G1 and G2) values.

Comprehensibility In Experiment 3, we analyze the clarity of each value and the distinguishability between value pairs.

Consistency. In Experiment 3, we analyze the consistency between Covid-G1 and Covid-G2, and Energy-G1 and Energy-G2 using crowdsourced opinion annotations.

4.1 Experiment 1: Value Lists

Four graduate students and two postdoctoral researchers, each working on a values-related research topic, participated as value annotators in Experiment 1. Two of these participants had a *technology and policy making* background, and four had a *computing* background. The two groups, G1 and G2, were constructed to have one member with *technology and policy making* background and two members with a *computing* background in each group.

4.1.1 *Opinion Corpora.*

We constructed two opinion corpora consisting of Dutch citizens' opinions in two different contexts using data collected via PVE surveys.

Covid Corpus contains opinions on *lifting Covid-19 measures in the Netherlands.* A PVE [25] for understanding participants' preferences on lifting Covid-19 related measures was conducted in the Netherlands between 29 April 2020 and 6 May 2020, when partial lockdown measures were in place in the Netherlands to limit the spread of Covid-19. The government had multiple plans for lifting such measures in the following weeks and months, and wanted to gauge Dutch citizens' opinions on the subject via PVE.

Energy Corpus contains opinions on *future energy policies for the Súdwest Fryslân municipality* in the Netherlands. The municipality's goal is to transition to renewable energy use, and there are multiple energy policies to achieve that goal. A PVE [40] was conducted between 10 April 2020 and 3 May 2020 to understand Súdwest Fryslân residents' opinions about the different energy policies.

The opinions in both Covid and Energy corpora were originally in Dutch. Since not all value annotators were fluent in Dutch, the opinions were translated to English using the MarianMT translator [17]. Further, opinions that contained only stop words or punctuation were removed. Then, the Covid corpus contained 59,461 and the Energy corpus contained 3,221 opinions.

4.2 Experiment 2: Context Specificity

Two graduate students with *technology and policy making* background participated in this experiment to evaluate the context-specificity of values. The two participants were familiar with the Covid and Energy contexts in which the PVEs were conducted. However, these two participants were not involved in Experiment 1; thus, they did not know which value belonged to which list.

We created a value list V_{CE} as the union of Covid-G1, Energy-G1, Covid-G2 and Energy-G2. Then, for each value $v \in V_{CE}$, we asked each participant the extent to which they agree with the following statement (once for Covid and once for Energy context) on a Likert scale of 1 (strongly disagree) to 5 (strongly agree).

> If I am a policy maker in the Covid (or Energy) context, knowing citizens' preferences about value v would help me in making a policy decision in that context.

We shuffled V_{CE} before asking the questions above so that each participant saw the values in a random order. For each value, we showed its name, keywords, and defining goal.

The two participants worked independently. After an initial round of ratings, the Intra-Class Correlation (ICC) between the two raters , an inter-rater reliability (IRR) metric for ordinal data [14], was 0.77. To ensure that the two participants had the same understanding of the task, they discussed their disagreements and performed another individual round of ratings. The ICC after the second round was 0.91, which is considered excellent [14].

4.3 Experiment 3: Comprehensibility and Consistency

To evaluate the comprehensibility of values in a list and the consistency between value lists for the same context, we employed 52 Prolific (www.prolific.co) crowd workers. The workers were directed to the Axies web application to complete evaluation.

Each crowd worker was assigned one value list. First, each worker was asked to read the information provided on the concept of values, and the context corresponding to the value list assigned to the worker. Then, each worker performed three tasks.

Clarity For each value in the list assigned to a worker, given the value name, keywords, and defining goal, the worker was asked to answer the following question on a five-point Likert scale.

> How clear do you find the concept described by the value above?

Distinguishability First, for a value list V, we computed the set P_V of all value pairs: $\forall v_i, v_j \in V : v_i \neq v_j, \{v_i, v_j\} \in P_V$. Then, we showed a subset of value pairs from P_V (along with the respective keywords and defining goals) to each worker assigned to the list V. For each value pair shown, the worker was asked to answer the following question on a five-point Likert scale.

> How distinguishable do you consider the two concepts to be?

Opinion annotation The final task for the crowd workers was to annotate opinions with values. First, we randomly selected 100 opinions from each opinion corpus. Then, we asked each worker assigned to a value list V to annotate a subset of the opinions selected for V's context. For each opinion, a worker could select one or more values from V, or mark the opinion as not value-laden.

We use the opinion annotations for evaluating the consistency of value lists. Since the same 100 opinions were annotated for both value lists for a context, we can measure the association between values in the two lists based on the opinions annotated with those values. For example, if the same set of opinions are annotated with $v_1 \in$ Covid-G1 and $v_2 \in$ Covid-G2, then we consider v_1 and v_2 as closely associated. Then, we (qualitatively) assess the consistency between Covid-G1 and Covid-G2 (similarly, Energy-G1 and Energy-G2) based on the the extent to which each value in Covid-G1 is associated with one or more values in Covid-G2.

Table 3 shows the number (#) of workers assigned to each value list, and the numbers of values, value pairs, and opinions assigned to each worker. The number of workers for each list was sufficient to obtain three annotations per opinion and three distinguishability ratings per value pair (one worker in each list annotated fewer than the shown number of pairs since that was sufficient to get three ratings per pair). Each worker rated all values in the assigned list.

4.3.1 Quality Control. The crowd workers were required to be fluent in English, and have submitted at least 100 tasks with at least 95% acceptance rate. We included four attention check questions: two in distinguishability rating and two in opinion annotation task.

Table 3: Overview of the crowd task

Value List	#Workers	#Values	#Value pairs	#Opinions
Covid-G1	12	11	14	25
Covid-G2	10	9	11	30
Energy-G1	15	14	19	20
Energy-G2	15	13	16	20

A total of 89 workers completed the task. We included a worker's task in our analysis only if the worker (1) passed both attention checks during distinguishability rating; and (2) at least one attention check during opinion annotation (we used one instead of two as the cutoff because there there was some room for subjectivity in answering the two attention check questions asked during opinion annotation). These criteria were set before any analysis of crowd work was done. Of the 89 workers, 52 satisfied the criteria above.

We suggested the time required for task completion (liberal estimate) as 45 minutes. The mean time spent by a crowd worker on our task was 30 minutes (with 16 minutes standard deviation). Each worker was paid £5.6 (at the rate of £7.5 per hour).

5 RESULTS AND DISCUSSION

We discuss the main results from our experiments in this section.

5.1 Value Lists

5.1.1 Exploration. A total of 12 explorations (six per context) were performed. In the Covid context, the mean time for exploration was 69.17 minutes (SD = 12.01 minutes) and the mean number of values annotated was 11.17 (SD = 2.64). In the Energy context, the mean time for exploration was 67.5 minutes (SD = 10.84 minutes) and the mean number of values annotated was 12.83 (SD = 5.23).

5.1.2 Consolidation. A total of four consolidations were performed (two groups of three annotators each; two consolidations, one per context, for each group), producing four value lists. The times spent in consolidating Covid-G1, Energy-G1, Covid-G2, and Energy-G2 were 105, 110, 115, and 120 minutes, respectively. After consolidations, the number of values per value list was Covid-G1: 11, Energy-G1: 14, Covid-G2: 9, and Energy-G2: 13.

5.2 Context Specificity

To evaluate the context specificity of a value list, we measure the extent to which the values in the list can influence policy decisions in the context for which the list was produced compared to a different context. We compute the specificity of a value v for a context c, as the mean of the ratings the two policy experts gave to value v for the context c. Recall that the policy experts were not aware of the context for which a value was annotated, a priori. The policy experts spent three hours each to rate the specificity of value lists.

Figure 6 (left) compares the specificity of Covid (including both Covid-G1 and Covid-G2) values for Covid and Energy contexts. Figure 6 (right) compares the specificity of Energy (including both Energy-G1 and Energy-G2) values for Covid and Energy contexts. We perform Wilcoxon's ransum test [15], a nonparametric test for ordinal data, to compare the two samples of ratings in each context and measure the effect size via Cliff's delta [7].

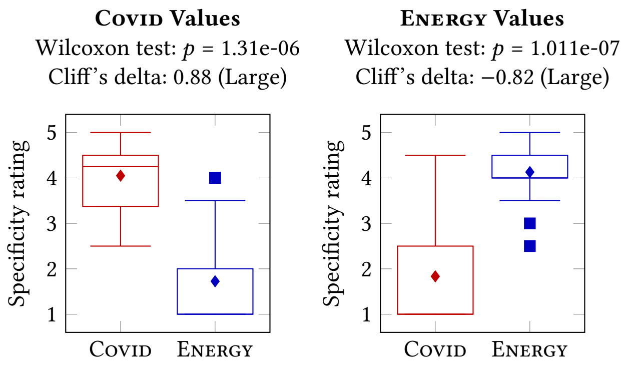


Figure 6: The context specificity of value lists

First, we observe that Covid values have significantly ($p < 0.05$) higher specificity ratings for the Covid context than the Energy context with a large effect size. Similarly, Energy values have significantly ($p < 0.05$) higher specificity ratings for the Energy context than the Covid context with a large effect size (the negative effect size is because of the ordering of the samples). This confirms our first hypothesis that Axies yields context-specific values.

Second, the specificity of a few values is low for their own contexts. Specifically, Care (Covid) and Representation (Energy) are rated less than 3 for their respective context. We observe that these two values are phrased broadly, and they may need refinement.

Finally, the specificity of some values were high for both contexts. Specifically, the Covid values of Autonomy and Equality were rated higher than 3 for the Energy context. Similarly, the Energy values of Inevitability, Distributional justice, Community, and Support were rated higher than 3 for the Covid context. Thus, some values can be applicable to more than one context.

5.3 Comprehensibility

We employ crowdsourced data to evaluate the clarity of values and the distinguishability between value pairs in a value list.

5.3.1 Clarity Evaluation. Recall that the clarity of a value in a list was rated by each crowd worker assigned to that list, yielding at least ten clarity ratings (Table 3) per value. Figure 7 shows the distribution of mean clarity ratings of Covid and Energy values.

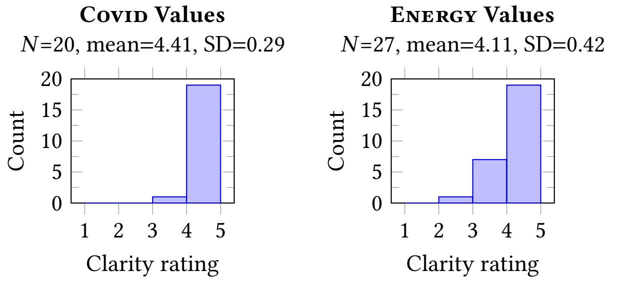


Figure 7: Histograms of value clarity ratings

Remarkably, the mean clarity rating of all but one value (among values in all four lists) was at least 3. Further, a large majority (80.9%) of the values received a mean clarity rating of at least 4. This suggests that Axies yields value lists clear to end users.

The Energy value of Distrust received the clarity rating of less than 3. The Distrust value has the defining goal "Big players (government, large companies) should not be in charge of solving problems on citizens' behalf." Perhaps, the connection between the Distrust value's name and its defining goal is not obvious, and we conjecture this as the reason for the value's low clarity rating.

Overall, the mean clarity rating of Covid values (4.41) was higher than that of Energy values (4.11). A potential reason for this is the timeliness of the Covid value lists; since people are currently experiencing the effects of the Covid-19 pandemic, they are able to better understand the value associated with the Covid context.

5.3.2 Distinguishability Evaluation. For each value pair in a value list, three crowd workers indicated how distinguishable the values in the pair were. Figure 8 shows the mean distinguishability ratings for pairs of values in the Covid and Energy value lists.

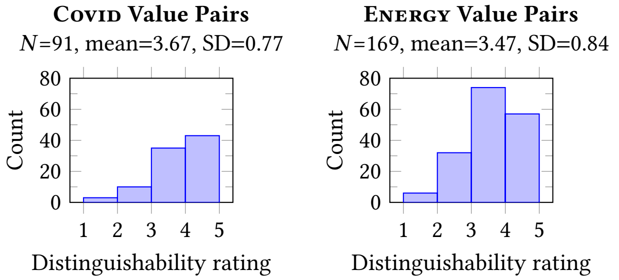


Figure 8: Histograms of value distinguishability ratings

None of the value pairs (among the four lists) have the mean distinguishability rating of 1. That is, no two value in any of the value lists are rated as indistinguishable. However, a good number of value pairs—14.3% Covid value pairs and 22.5% Energy value pairs—have a mean distinguishability rating in (1, 3). Thus, although Axies yields distinct values for a context, the values in a context have similarities among them. This observation aligns with Schwartz's [35] postulate that values form a continuum of related motivations.

5.4 Consistency

To evaluate the consistency between the two value lists for the same context, we employ the crowdsourced opinion annotations. For example, let $v_1 \in$ Covid-G1 and $v_2 \in$ Covid-G2, and O_1 and O_2 be the set of opinions annotated with v_1 and v_2, respectively. We consider an opinion o as annotated with a value v if at least two of the three annotations for o include v. Then, we measure the association between the two values as the Jaccard similarity between their opinion annotations:

$$J(v_1, v_2) = \frac{|O_1 \cap O_2|}{|O_1 \cup O_2|} \tag{2}$$

For each value in one value list for a context, Figure 9 shows the closest value in the other list for the context, to emphasize the associations between the two lists.

Although value lists for the same context are not identical, we observe that each value in one list for a context is associated (has

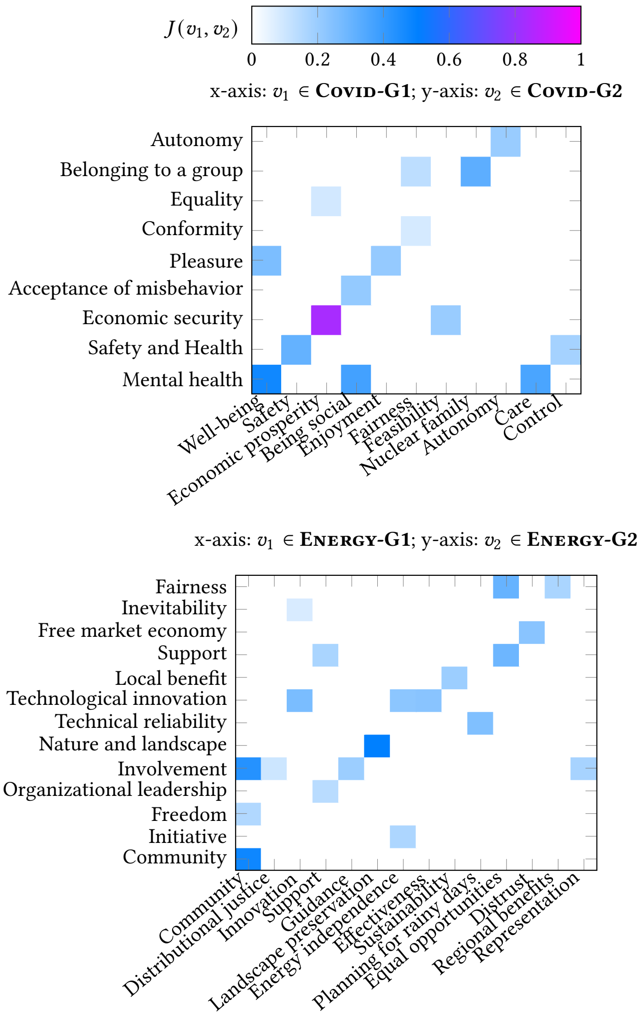


Figure 9: Association between G1 and G2 value lists

a non-zero Jaccard similarity) with at least one value in the other list for that context. In some cases, the association is apparent from the value names, e.g., Economic prosperity ∈ Covid-G1 and Economic security ∈ Covid-G2. In some cases, despite differences in the names, the values capture similar motivations, e.g., Planning for rainy days ∈ Energy-G1 and Technical reliability ∈ Energy-G2, capture the same motivational goal of planning for unforeseen circumstances. In some cases, the motivation behind a value in a list was distributed over more than one value in the other list. For example, Fairness ∈ Energy-G2 is captured by Equal opportunities and Regional benefits ∈ Energy-G1. In essence, no value is conceptually exclusive to one value list within a context.

6 CONCLUSIONS AND DIRECTIONS

Axies combines human and artificial intelligence to yield context-specific values. In a specific context, e.g. driving, context-specific values can be more effective in explaining and predicting human behavior than general values [44]. An autonomous driving agent can concretely elicit its passengers' preferences over driving-specific values (e.g., safety and efficiency) in order to tailor the driving experience. However, the tradeoffs between context-specific and general values must be empirically investigated over multiple context. Axies provides the key ingredient for such investigations: a list of context-specific values to compare the general values with.

Our experiments highlight important properties of Axies. First, Axies yields values that are clear to the end users. The clarity is important for an agent to (1) elicit value preferences from users, e.g., by asking whether Mental health is more important to a user than Conformity in a context, and (2) explain that the agent made a certain decision because the agent inferred, e.g., Fairness as more important to the user than Regional benefits in the decision context.

Second, we find that Axies yields distinct values for a context. However, based on value annotators feedback and crowd distinguishability results, we observe that values in a context have similarities since they form a motivational continuum. An interesting research direction is to identify and visualize a value continuum (e.g., as a circumplex [35]) from a list of context-specific values. We cobjecture that such a visualization would support annotators in the process of building a cohesive value list.

Third, as a methodology, we expect Axies to yield reproducible results. Following Axies to annotate an opinion corpus should yield consistent value lists independent of the annotators. However, considering the subjective human judgements involved, we do not expect a value list produced for a context by one group to be identical to the value list produced by another group. As expected, the value lists generated for the same context by different groups of annotators are not identical but consistent in that each value in one list is associated with one of more values in the other list.

Fourth, a key result from our experiments is that Axies yields context-specific values as it set out to. Specifically, we observe that the values identified for a context are more useful for decision making in that context than another context. However, some context-specific values are more broadly applicable than others.

Identifying context-specific values is a significant effort. Axies simplifies this process and systematically guides the annotators, who need not be design experts. An interesting future direction is to analyze the benefits of NLP and active learning on the overall process (e.g., by comparing Axies to a baseline without the AI components). Further, in our experiments, the annotators followed the Axies steps one time. In practice, Axies can be used in an agile manner with multiple exploration-consolidation sprints with feedback from evaluations in between the sprints.

Context-specific values must be easy to discover to support reasoning and analyses. To this end, a repository of context-specific values, where values are linked with contexts, opinions, and application scenarios would be valuable. Given such a repository, designers and developers can reuse values suitable for their contexts, and an agent can automatically pick relevant values for a decision context.

ACKNOWLEDGMENTS

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research.

REFERENCES

[1] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. IFAAMAS, Auckland, New Zealand, 16–24.

[2] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn J. M. Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (2020), 18–28.

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

[4] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2019. Incorporating Behavioral Constraints in Online AI Systems. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI '19)*. AAAI Press, Honolulu, Hawaii, USA, 3–11.

[5] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active Semi-Supervision for Pairwise Constrained Clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM '04)*. Society for Industrial and Applied Mathematics, Orlando, Florida, USA, 333–344.

[6] Ryan L. Boyd, Steven R. Wilson, James W. Pennebaker, Michal Kosinski, David J. Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the 9th International Conference on Web and Social Media (ICWSM '15)*. AAAI Press, Oxford, UK, 31–40.

[7] Norman Cliff. 2014. *Ordinal methods for behavioral data analysis*. Psychology Press, Hove, East Sussex, UK.

[8] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI '17)*. AAAI Press, San Francisco, California, USA, 4831–4835.

[9] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 178–184.

[10] Batya Friedman, Peter H. Kahn, and Alan Borning. 2008. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 69–101.

[11] Justin Garten, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods* 50, 1 (2018), 344–361.

[12] Barney G. Glaser and Anselm L. Strauss. 1967. *The discovery of grounded theory*. Aldine Publishing, Chicago, Illinois, USA.

[13] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology* 96, 5 (2009), 1029–1046.

[14] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology* 8, 1 (2012), 23–34.

[15] Myles Hollander and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. Wiley, New York, USA.

[16] Joe Hoover, Kate Johnson, Reihane Boghrati, Jesse Graham, and Morteza Dehghani. 2018. Moral framing and charitable donation: integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology* 4, 1 (2018), 1–18.

[17] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations (ACL '18)*. Association for Computational Linguistics, Melbourne, Australia, 116–121.

[18] Christopher A. Le Dantec, Erika Shehan Poole, and Susan P. Wyche. 2009. Values as lived experience. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI '09)*. ACM Press, New York, USA, 1141.

[19] Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring Background Knowledge to Improve Moral Value Prediction. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '18)*. IEEE, Barcelona, 552–559.

[20] Enrico Liscio, Michiel van der Meer, Catholijn M. Jonker, and Pradeep Murukannaiah. 2021. Axies: Identifying and Evaluating Context Specific Values - code. https://doi.org/10.4121/13712908

[21] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context Specific Values - supplemental material. https://doi.org/10.4121/13705423

[22] Hui Liu, Yinghui Huang, Zichao Wang, Kai Liu, Xiangen Hu, and Weijun Wang. 2019. Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System. *Applied Sciences* 9, 10 (2019), 1992.

[23] Rijk Mercuur, Virginia Dignum, and Catholijn M. Jonker. 2019. The value of values and norms in social simulation. *Journal of Artificial Societies and Social Simulation* 22, 1 (2019), 9.

[24] Niek Mouter, Paul Koster, and Thijs Dekker. 2021. Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. *Transportation Research Part A: Policy and Practice* 144 (2021), 54 – 73.

[25] Niek Mouter, Shannon L. Spruit, Anatol V. Itten, José Ignacio Hernandez, Lisa Volberda, and Sjoerd Jenninga. 2020. Leaving the smart lockdown together: results of consulting 30,000 Dutch citizens on relaxing corona measures. www.tudelft.nl/covixexit

[26] Nikola Mrkšić, Diarmuid Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei Hao Su, David Vandyke, Tsung Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '16)*. Association for Computational Linguistics, San Diego, California, USA, 142–148.

[27] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn J. M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. IFMAAMAS, Aukland, New Zealand, 1706–1710.

[28] Pradeep K. Murukannaiah and Munindar P. Singh. 2014. Xipho: Extending tropos to engineer context-aware personal agents. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '14)*. IFAAMAS, Paris, France, 309–316.

[29] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn M. Jonker. 2012. Elicitation of situated values: Need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology* 14, 4 (2012), 285–303.

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP '19)*. Association for Computational Linguistics, Hong Kong, China, 3973–3983.

[31] Milton Rokeach. 1973. *The nature of human values*. Free Press, New York, USA.

[32] Daniel J. Rosenkrantz, Richard E. Stearns, and Philip M. Lewis II. 1977. An analysis of several heuristics for the traveling salesman problem. *SIAM J. Comput.* 6, 3 (1977), 563–581.

[33] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Magazine* 36, 4 (2015), 105–114.

[34] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality and Quantity* 52, 4 (2018), 1893–1907.

[35] Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture* 2, 1 (2012), 1–20.

[36] Marc Serramia, Maite Lopez-Sanchez, and Juan A. Rodriguez-Aguilar. 2020. A Qualitative Approach to Composing Value-Aligned Norm Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. IFAAMAS, Auckland, New Zealand, 1233–1241.

[37] Nate Soares. 2014. *The Value Learning Problem.* Technical Report. Machine Intelligence Research Institute, Berkeley, California, USA. 8 pages.

[38] Nate Soares and Benya Fallenstein. 2017. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity: Managing the Journey*. Springer, Berlin, 103–125.

[39] Daniel J. Solove. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154, 3 (2006), 477–560.

[40] Shannon L. Spruit and Niek Mouter. 2020. 1376 residents of Súdwest-Fryslân about the future energy policy of their municipality: the results of a consultation. https://www.tudelft.nl/en/tpm/pve/case-studies/energy-in-sudwest-fryslan/

[41] Livia Teernstra, Peter van der Putten, Liesbeth Noordegraaf-Eelens, and Fons Verbeek. 2016. The Morality Machine: Tracking Moral Values in Tweets. In *Advances in Intelligent Data Analysis XV: 15th International Symposium (IDA '16)*. Springer, Stockholm, Sweden, 26–37.

[42] Andrea Aler Tubella and Virginia Dignum. 2019. The glass box approach: Verifying contextual adherence to values. In *Proceedings of the Workshop on Artificial Intelligence Safety (AISafety '19)*. CEUR–WS, Macao, China, 68–74.

[43] Ibo van de Poel. 2013. Translating Values into Design Requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*. Springer Netherlands, Dordrecht, Netherlands, 253–266.

[44] Tom G.C. van den Berg, Maarten Kroesen, and Caspar G. Chorus. 2020. Does morality predict aggressive driving? A conceptual analysis and exploratory empirical investigation. *Transportation Research Part F: Traffic Psychology and Behaviour* 74, 1 (2020), 259–271.

[45] W. Fred van Raaij and Theo M. M. Verhallen. 1994. Domain-specific Market Segmentation. *European Journal of Marketing* 28, 10 (1994), 49–66.

[46] Steven R. Wilson, Yiting Shen, and Rada Mihalcea. 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)*. Springer, St. Petersburg, Russia, 455–470.