**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.vander.meer@lia.cs.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:**


**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 αξίας; 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 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 context by exploring opinions about the context, and (2) think about values one opinion at a time.

Further, Axies distributes this task among a small group of annotators. In 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.


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’ response to 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).
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

<table>
<thead>
<tr>
<th>Preference</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nursing homes allow visitors again</td>
<td>Loneliness and isolation are a bigger killer than Corona. Someone’s got to keep the economy going.</td>
</tr>
<tr>
<td>All restrictions are lifted for persons who are immune</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Keywords</th>
<th>Defining goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental health</td>
<td>Loneliness, quality of life, stress</td>
<td>The strive towards protecting and improving one’s emotional and psychological well-being.</td>
</tr>
<tr>
<td>Economic prosperity</td>
<td>Economy, stability, bankruptcy</td>
<td>Being able to pay and afford what you need.</td>
</tr>
</tbody>
</table>

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 $n_o$ be the name and $K_o = \{k_{o1}, \ldots, k_{oK}\}$ be the set of keywords of a value $o$. Then, Axies computes the value vector $M(v)$ using the Distributed Dictionary Representation [11] as:

$$M(v) = \frac{M(n_o) + \sum_{k \in K_o} M(k)}{|M(n_o) + \sum_{k \in K_o} M(k)|}$$

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.
Distance between Values in the Pair

FFT Distance

We suggest the annotators to reach thinking whereas consolidation facilitates convergent thinking.

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.

### Algorithm 1: Fetching next opinion using FFT

| Input: \( O, M \); /* Opinions, Embedding model */ |
| Output: \( V_\alpha \); /* Value list of a */ |
| initialization: \( \forall o \in O: d_o = \infty; V_\alpha = \emptyset; \) |
| while \( O \neq \emptyset \) && ~saturated\( (V_\alpha) \) do |
| \( o_{\text{next}} = \arg \max_{o \in O} d_o; /* \text{break ties randomly */} \) |
| \( O = O - o_{\text{next}}; \) |
| \( V_\alpha^\text{old} = V_\alpha; \) |
| update_values\( (V_\alpha, o_{\text{next}}) \); |
| \( V_\alpha^\delta = V_\alpha - V_\alpha^\text{old}; \) |
| \( \forall o \in O: d_o = \) |
| \( \min \{ \cosine\_\text{distance}(M(o), M(o_{\text{next}})), \} \); |
| \( \forall o \in V_\alpha^\delta: \cosine\_\text{distance}(M(o), M(z)) \}; \) |
| 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.

#### 3.5.1 Procedure

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

#### 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 \( V_\alpha \), and repeat that process several times, to arrive at a refined \( V_\alpha \).

To reduce the cognitive load, Axies controls the order in which value pairs are presented to the annotators—the most similar value pair from \( V_\alpha \) is shown first. This approach is beneficial because similar values are likely to be merged, reducing the size of \( V_\alpha \), 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 \( V_\alpha \).

#### 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 3: Example progress plot of exploration](image)

![Figure 4: Example progress plot of consolidation](image)
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](image)

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 Südwest 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 Südwest 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 agreed 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 and values. For example, if the same set of opinions are annotated as values in the two lists based on the opinions annotated with those value lists for a context, we can measure the association between value name, keywords, and defining goal, the worker was asked to show 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:

**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 $v_1 \in \text{COVID-G1}$ and $v_2 \in \text{COVID-G2}$, then we consider $v_1$ and $v_2$ as closely associated. Then, (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.

<table>
<thead>
<tr>
<th>Value List</th>
<th>#Workers</th>
<th>#Values</th>
<th>#Value pairs</th>
<th>#Opinions</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-G1</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>COVID-G2</td>
<td>10</td>
<td>9</td>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td>Energy-G1</td>
<td>15</td>
<td>14</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Energy-G2</td>
<td>15</td>
<td>13</td>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>

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].
We employ crowdsourced data to evaluate the clarity of values and 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.

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, Distributitional 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.

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.

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 \(v_1 \in \text{COVID-G1} \) and \(v_2 \in \text{COVID-G2}\), 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:

\[
J(v_1, v_2) = \frac{|O_1 \cap O_2|}{|O_1 \cup O_2|}
\]

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
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 contexts. 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 conjecture 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


