A collaborative platform for identifying context-specific values

Liscio, E.; Meer. M.T. van der; Jonger, C.M.; Murukannaiah, P.K.

A Collaborative Platform for Identifying Context-Specific Values

Demonstration Track

Enrico Liscio
E.Liscio@tudelft.nl
Delft University of Technology, the Netherlands

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

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

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

ABSTRACT

Value alignment is a crucial aspect of ethical multiagent systems. An important step toward value alignment is identifying values specific to an application context. However, identifying context-specific values is complex and cognitively demanding. To support this process, we develop a methodology and a collaborative web platform that employs AI techniques. We describe this platform, highlighting its intuitive design and implementation.

KEYWORDS

Values; Ethics; Context; Natural Language Processing

ACM Reference Format:
Enrico Liscio, Michiel van der Meer, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2021. A Collaborative Platform for Identifying Context-Specific Values: Demonstration Track. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

Values are abstract motivations that guide our opinions and actions [15]. Engineering value-sensitive agents that learn and align their actions with human values is essential for robust and beneficial artificial intelligence (AI) [3, 10, 11, 14, 17]. Then, an important question is: what values should an agent learn and align with?

Several lists of *basic values*, that transcend cultures and contexts, have been described in the literature [4, 6, 15]. However, a growing number of researchers emphasize that values must be situated within an application context for concrete analysis, e.g., to reason about conflicting values [1, 12], align values and norms [16], or evaluate value adherence of an agent-based system [18].

We define *context-specific values* as values "applicable and defined specifically within a context" [9]. The following scenario illustrates why context-specific values are important for an agent. Consider a personal travel agent. Schwartz values [15] of security and hedonism are relevant to the agent's reasoning but the value of power is arguably not. Further, to ensure security, the agent aims at increasing travel safety. However, travel safety takes different meanings in different contexts: during a pandemic, it is safer to travel by car to avoid larger crowds; otherwise, traveling by public transportation is preferable to reduce the likelihood of accidents.

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.

As context-specific values vary with contexts, we need an efficient and reusable approach to identify context-specific values. We propose Axies [9], a methodology to systematically identify context-specific values. Axies has two key features: (1) it requires collaborative work among human annotators, who perform several high-level cognitive tasks, and (2) it exploits natural language processing (NLP) and active learning techniques to guide annotation.

Axies is a hybrid methodology in that human annotators are supported by AI in the process of value identification. However, Axies annotators (e.g., citizens and policy makers) may not have AI expertise. Further, Axies requires collaboration among the annotators. Thus, a computational platform is necessary to support the annotators in applying Axies without exposing them to the underlying technical mechanisms. To enable these features, we develop an intuitive and reusable web platform[1], with an AI back-end, on which human annotators can collaborate. In this paper, we describe the design and implementation of this platform.

2 AXIES PLATFORM

Identifying the values relevant to a context is challenging. To simplify this task, Axies employs AI techniques to guide a small group of *annotators* through an opinion corpus composed of value-laden textual *opinions* about a context. Axies promotes *inductive reasoning* by asking annotators to annotate values based on the opinions. A value is described by its *name*, *keywords* (words that help binding the value to the context) and *defining goal* (which describes what holding a value in the context means). The result of Axies is a *value list* relevant to the authors of the opinions in the examined context.

The Axies methodology is composed of two phases: an individual value annotation phase (*exploration*) and a collaborative merge of the individual value lists (*consolidation*). Our web platform supports both phases as described in the following subsections.

2.1 Implementation Details

The platform is implemented in Python on the Flask micro web framework [7]. The back end is also implemented in Python to provide seamless integration with state-of-the-art NLP models. All data is stored in a SQLite database [8]. Further, we developed functionalities to import the opinion corpus in a `csv` or `yaml` format. Finally, the responsive web interface is implemented in JavaScript. The interface can be used on small (e.g., smart phone) and large screens, and it utilizes the de facto standards in modern web applications.

[1]Demonstration: https://youtu.be/s7nJPr2Z80w

The modular setup of the two phases enables easy extension to new annotation tasks. The source code is available on GitHub[2].

2.2 User Navigation

Annotators are required to register with a username and a password. Operations can be performed asynchronously. Data is stored to the SQL database upon input, allowing the annotators to leave and return to the platform without losing progress.

A top navigation bar is accessible from any page (as shown in Figure 1), permitting users to switch between the two phases of Axies (`Explore` and `Consolidate`) and different contexts (e.g., Covid and Energy in the case of the experiments in [9]).

2.3 Exploration

During the exploration phase, annotators individually generate a value list based on the opinions in the corpus. However, opinion corpora may be too large to be analysed by an individual. Axies aims at exposing the annotators to a subset of the corpus while increasing the coverage of read opinions. Active learning and NLP techniques support the exploration phase by controlling the order in which annotators are presented with the opinions in the corpus.

The web platform reduces information overload by presenting one opinion at a time for annotation as shown at the top of Figure 1. Annotators are asked to annotate values and keywords based on the shown opinion. The interface allows them to add and delete values and keywords at any moment.

To select an opinion for annotation, first, all opinions are encoded to a vector space through the sentence embedding Sentence-BERT model [13]. Distributed Dictionary Representation [5] allows encoding values to the same embedding space. Then, the Farthest First Traversal [2] algorithm selects the next opinion to be annotated as the farthest in the embedding space from the values already annotated and the opinions already shown to the annotators.

The *progress plot* (on the right of Figure 1) contains a bar per each opinion shown to an annotator, where the color indicates the actions (or lack thereof) performed upon reading the opinion. This intuitive visualization assists annotators in keeping track of their progress and deciding when saturation is reached. Finally, each value is associated with a button to fetch opinions similar to the value in order to refine individual value concepts.

2.4 Consolidation

During the consolidation phase, annotators are invited to combine their individual value lists. While exploration promotes divergent thinking, consolidation promotes convergent thinking. To simplify consolidation, Axies creates the union of all individual value lists and guides the annotators in methodically refining it. To facilitate this process, annotators are sequentially presented with just a pair of values at a time. Axies selects the pair as the most similar values in the vector space, assuming them as the most likely to be merged.

For each value in the pair, the annotators can fetch the opinions that led to the value annotation during exploration. If the annotators deem the two values to be conceptually identical, they may merge them by using the interface offered by the platform. Alternatively, the may edit the values in the pair and the whole value list at any moment. Upon consolidation of the value pair, annotators may fetch the following pair suggested by Axies, or decide to manually fetch the next pair from the value list. As in the exploration phase, a progress plot helps the annotators in tracking their progress. Finally, when consolidation of the list is terminated, annotators are asked to add a defining goal to each value.

[2]Code: https://github.com/enricoliscio/axies

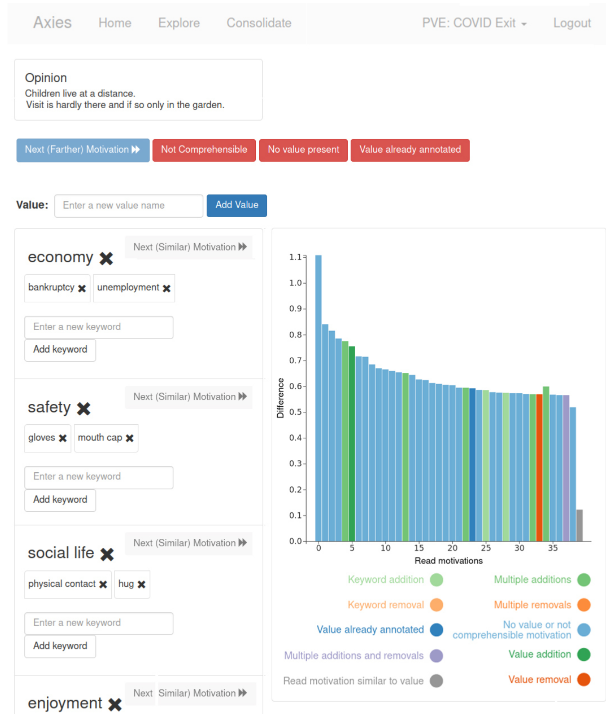

Figure 1: Exploration in the web application

3 CONCLUSION

We present the Axies platform, which simplifies the complex value identification task as a guided value annotation task. Our platform successfully supported the experiments involving two contexts and two groups of annotators [9] by providing an intuitive design that allows the annotators to visualize all components. The experiments show that Axies yields values that are context-specific, comprehensible to laypeople and consistent across different annotators.

Based on the feedback received by the participants in our experiments, we identify three main directions for future work. First, developing techniques to visualize values by highlighting their similarities and differences can help annotators in generating more comprehensive value lists. Second, during consolidation annotators often examined the proposed value pairs without taking actions, and sometimes resorted to selecting value pairs manually. The consolidation phase would benefit of an improved value pair selection, e.g., by normalizing the impact of keywords when computing value embeddings. Finally, as value lists emerge for multiple contexts, we call for maintaining an open-access repository of values and associated contexts. Such a repository would enable researchers in studying connections among value lists, and designers and developers in choosing values suitable for their applications.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Miguel Grinberg. 2018. *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc., Boston, Massachusetts, USA.

[8] Richard D. Hipp. 2020. SQLite. https://www.sqlite.org/index.html

[9] 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 *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '21)*. IFAAMAS, Online, 1–10.

[10] 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.

[11] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.