Identifying Context-Specific Values via Hybrid Intelligence

Liscio, Enrico; Jonker, Catholijn M.; Murukannaiah, Pradeep K.

DOI
10.3233/FAIA220223

Publication date
2022

Document Version
Final published version

Published in
HHAI2022

Citation (APA)
Liscio, E., Jonker, C. M., & Murukannaiah, P. K. (2022). Identifying Context-Specific Values via Hybrid Intelligence. In S. Schlobach, M. Perez-Ortiz, & M. Tielman (Eds.), *HHAI2022: Augmenting Human Intellect - Proceedings of the 1st International Conference on Hybrid Human-Artificial Intelligence* (pp. 298-301). (Frontiers in Artificial Intelligence and Applications; Vol. 354). IOS Press. https://doi.org/10.3233/FAIA220223

HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220223

Identifying Context-Specific Values via Hybrid Intelligence

Enrico Liscio [a,1], Catholijn M. Jonker [a,b] and Pradeep K. Murukannaiah [a]
[a] *Interactive Intelligence (II), Delft University of Technology*
[b] *Leiden Institute for Advanced Computer Science (LIACS), Leiden University*

1. Introduction

Values, e.g., benevolence and self-determination, are abstract motivations that explain and justify human behavior and opinions [19]. Values are instrumental for hybrid intelligence (HI) systems [1, 14, 18, 20, 21] that involve humans and artificial agents. Then, a crucial question to be answered is: what values should an artificial agent align with?

Lists of *general values*, which are applicable across cultures and contexts, have been crafted by ethicists [17, 19], political scientists [6], and designers (e.g., Value Sensitive Design [5]). For example, the Schwartz value list [19], a highly influential list of general values [7], includes values such as self-direction, power, and security. However, context is a crucial factor when reasoning about values. (1) Not all values are relevant to all contexts [8, 15, 19]. (2) The way in which we express value rhetoric differs from one context to another [11]. (3) Preferences over general values may not be consistent across contexts [4] – that is, our interpretation and prioritization of values is influenced by context.

General values help explain broad human behavioral tendencies, such as attitude toward immigration and activism [3]. However, for concrete applications, values must be situated within a context. Thus, we define *context-specific values* as values applicable and defined within a context. Consider, for example, the task of value elicitation [8] – identifying individuals' preferences over competing values – for the intent of decision-making on green energy transition. For this concrete task, we can elicit stakeholders' preferences between two context-specific values such as landscape preservation and energy independence, or between two general values such as security and self-direction. We expect that choosing between the context-specific values is easier for laypeople to justify and more insightful for policy makers than choosing between the general values.

Contribution In this extended abstract, we summarize our work previously published at AAMAS and JAAMAS [9, 10, 12]. Our contribution in these papers is two-fold. (1) We propose Axies, a hybrid methodology for identifying context-specific values. (2) We evaluate Axies in a user study involving 80 human subjects. We compare Axies value lists generated for two contexts to the Schwartz (general) value list (due to its high contemporary influence [7]) in their context specificity, comprehensibility, consistency, and application. We also explore the relation between context-specific and general value lists.

[1] Corresponding Author: Enrico Liscio, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands; E-mail: e.liscio@tudelft.nl.

2. Axies Methodology

Axies is a hybrid (human + AI) methodology where Natural Language Processing (NLP) techniques support a small group of annotators in identifying values relevant to a context. The input to Axies is an opinion corpus, which includes users' *value-laden* opinions within a context. The output is a list of values relevant to the context under analysis. Axies stimulates *inductive reasoning* by inviting the annotators to identify values held by users based on the opinions express by the users themselves. A crucial advantage of this approach is that the resulting Axies values are grounded in data.

Axies is composed of two phases: exploration and consolidation. During *exploration*, annotators are independently guided through the opinion corpus to develop an individual value list. To do so, all opinions present in the corpus are represented as vectors using the Sentence-BERT (S-BERT, [16]) sentence embedding model. Annotators are exposed to one opinion at a time, selected as the farthest opinion from the already visited opinions through the Farthest First Traversal algorithm [2]. Upon reading an opinion, annotators are asked to write down the value(s) underlying the opinion (if present).

During *consolidation*, the annotators in a group collaborate to merge their individual value lists. All values present in the individual value lists are embedded with the S-BERT model. The two most similar values are presented to the annotators, who are asked to discuss and decide whether the two value concepts overlap and thus merge the values, or continue with the following value pair. The final result is a consolidated group value list.

3. Results and Discussion

In our experiments, we asked two groups of three annotators each to perform Axies on two opinion corpora. In each group, one annotator had a technology and policy making background, and two had a computer science background. The opinion corpora were composed of the answers to two surveys conducted on COVID-19 regulations [13] and green energy transition [22]. Examples of resulting values are `mental health` (COVID-19) and `landscape preservation` (energy). We then asked two policy-making experts and 72 crowd workers to evaluate the Axies value lists and compare them to the Schwartz value list, reaching the following five conclusions. (1) Axies yields *consistent* value lists for a context, independent of the annotators. (2) Laypeople deem Axies values *comprehensible* (that is, easy to understand and distinguishable one from another). (3) Values yielded by Axies for a context are more *specific* for that context than general values. (4) When put to the concrete *application* of value annotation, laypeople annotate Axies values more often and with higher agreement than general values. (5) Only a few general values have a clear correspondence to Axies values (i.e., only the general values that are relevant to the context), and general values with a clear correspondence are often related to multiple Axies values that describe them in a more fine-grained manner in the context.

Value alignment is recognized as a research priority for achieving beneficial AI [18]. Identifying the relevant values that an artificial agent ought to align with is a remarkable effort. Axies facilitates this process by employing NLP techniques to guide human annotators through a value-laden corpus. This hybrid nature allows annotators to minimize their effort and focus on few high-level actions. A compelling future direction is to investigate the benefits of the AI component on the value identification process (e.g., by comparing Axies to a fully manual baseline).

References

[1] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. J. M. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, and M. Welling. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[2] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, SDM '04, pages 333–344, Orlando, Florida, USA, 2004. Society for Industrial and Applied Mathematics. .

[3] G. Datler, W. Jagodzinski, and P. Schmidt. Two theories on the test bench: Internal and external validity of the theories of Ronald Inglehart and Shalom Schwartz. *Social Science Research*, 42(3):906–925, 2013. .

[4] J. de Wet, D. Wetzelhütter, and J. Bacher. Revisiting the trans-situationality of values in Schwartz's Portrait Values Questionnaire. *Quality and Quantity*, 53(2): 685–711, 2018. .

[5] B. Friedman, P. H. Kahn, and A. Borning. Value sensitive design and information systems. In *The Handbook of Information and Computer Ethics*, pages 69–101. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2008. .

[6] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier, Amsterdam, the Netherlands, 2013. .

[7] P. H. Hanel, L. F. Litzellachner, and G. R. Maio. An empirical comparison of human value models. *Frontiers in Psychology*, 9(SEP):1–14, 2018. .

[8] C. A. Le Dantec, E. S. Poole, and S. P. Wyche. Values as lived experience. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, page 1141, New York, USA, 2009. ACM Press. .

[9] E. Liscio, M. van der Meer, C. M. Jonker, and P. K. Murukannaiah. A Collaborative Platform for Identifying Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 1773–1775, Online, 2021. IFAAMAS.

[10] E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, N. Mouter, and P. K. Murukannaiah. Axies: Identifying and Evaluating Context-Specific Values. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 799–808, Online, 2021. IFAAMAS.

[11] E. Liscio, A. E. Dondera, A. Geadau, C. M. Jonker, and P. K. Murukannaiah. Cross-Domain Classification of Moral Values. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '22, pages 1–12, Seattle, USA, 2022. Association for Computational Linguistics. To appear.

[12] E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. What values should an agent align with? *Autonomous Agents and Multi-Agent Systems*, 36(23):32, 2022. .

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

[14] P. K. Murukannaiah, N. Ajmeri, C. J. M. Jonker, and M. P. Singh. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1706–1710, Aukland, New Zealand, 2020. IFAAMAS.

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

[16] N. Reimers and I. Gurevych. 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, pages 3973–3983, Hong Kong, China, 2019. Association for Computational Linguistics. .

[17] M. Rokeach. *The Nature of Human Values*. Free Press, New York, USA, 1973.

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

[19] S. H. Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):1–20, 2012.

[20] L. C. Siebert, E. Liscio, P. K. Murukannaiah, L. Kaptein, S. L. Spruit, J. van den Hoven, and C. M. Jonker. Estimating Value Preferences in a Hybrid Participatory System. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence*, HHAI '22, pages 1–14, Amsterdam, the Netherlands, 2022. IOS Press.

[21] N. Soares and B. Fallenstein. Agent foundations for aligning machine intelligence with human interests: A technical research agenda. In *The Technological Singularity: Managing the Journey*, pages 103–125. Springer, Berlin, 2017. .

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