HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230102

Inferring Values via Hybrid Intelligence

Enrico LISCIO [a,1], Roger LERA-LERI [b], Filippo BISTAFFA [b], Roel I.J. DOBBE [a], Catholijn M. JONKER [a,d], Maite LOPEZ-SANCHEZ [c], Juan A. RODRIGUEZ-AGUILAR [b], and Pradeep K. MURUKANNAIAH [a]

[a] *Delft University of Technology, Delft, the Netherlands*
[b] *IIIA-CSIC, Barcelona, Spain*
[c] *University of Barcelona, Barcelona, Spain*
[d] *Leiden University, Leiden, the Netherlands*

Abstract. Values, such as freedom and safety, are the core motivations that guide us humans. A prerequisite for creating value-aligned multiagent systems that involve humans and artificial agents is value inference, the process of identifying values and reasoning about human value preferences. We introduce a framework that connects the value inference steps, and motivate why a hybrid intelligence approach is instrumental for its success. We also highlight the multidisciplinary research challenges that hybrid value inference entails.

Keywords. values, norms, ethics, sociotechnical systems, hybrid intelligence

1. Introduction

Values, e.g., freedom and safety, are the core motivations that guide humans. The relative importance that an individual ascribes to different values (our *value preferences*) drives actions [32]. Values are crucial for sociotechnical systems (STS) [28] that involve humans and artificial agents. A prerequisite for creating a value-aligned STS is *value inference*, the process of identifying values and reasoning about stakeholders' value preferences [25]. However, since value reasoning is cognitively challenging [19, 29] and implicit in human thinking [16, 21], value inference cannot be performed solely via computational methods. A hybrid intelligence (HI) [1] approach is necessary to guide humans to become aware of their value preferences and how they change based on context.

In this extended abstract, we summarize a framework [25] that connects the value inference steps, and motivate why an HI approach is instrumental for its success. We also highlight the multidisciplinary research challenges that hybrid value inference entails.

2. Hybrid Value Inference

We propose a framework for hybrid value inference (Figure 1), composed of the steps to go from behavioral data to aggregated value preferences. As *behavioral data*, we consider stakeholders' actions (e.g., how they choose over competing alternatives [6, 35])

[1]Corresponding Author: Enrico Liscio, Delft University of Technology, the Netherlands, e.liscio@tudelft.nl

and justifications provided for those actions (since value preferences are often implicit in actions and language is our preferred way of expressing values [13, 31]). Then, *value identification* is the process of identifying the set of values relevant to a decision context. Inspired by Value Sensitive Design [11], we advocate for methods that take into account the decision context [22, 24] and involve stakeholders, e.g., via data-driven methods [8, 40, 41]. Subsequently, *value estimation* is the process of determining an individual's value preferences over the identified values. As language is our preferred way to express values, we envision value estimation to be based on both actions and justifications provided in a decision context (e.g., [35]), with the support of natural language processing methods (e.g., [3, 17, 23]). Finally, *value aggregation* is the process of aggregating individual value preferences into a societal value system. We encourage the use of computational social choice approaches (e.g., [12, 20]) that consider multiple consensuses and ethical principles at the same time, constructed interactively via explanations [7].

However, a sequence of computational methods applied on behavioral data is not likely to yield good estimates of individual and societal value preferences, as value preferences are often implicit to humans [16, 21, 39] and are, thus, not easily observable in behavioral data. Hence, we must actively engage humans via HI methods [1], requiring human and artificial intelligence to augment each other. On the one hand, humans must be made aware of values and guided through value reasoning via a process of *self-reflection* [21, 29]. Agents can facilitate self-reflection by situating value reasoning in specific contexts and behaviors, e.g., by asking concrete questions such as what motivated a human to choose a specific action in a decision context. On the other hand, *deliberation* with others [9] and confronting individuals with different value systems [30] help us in discovering our own value systems. To this end, an increasing number of digital deliberation platforms have been proposed [18, 34], where artificial moderating agents can facilitate large-scale deliberation [15].

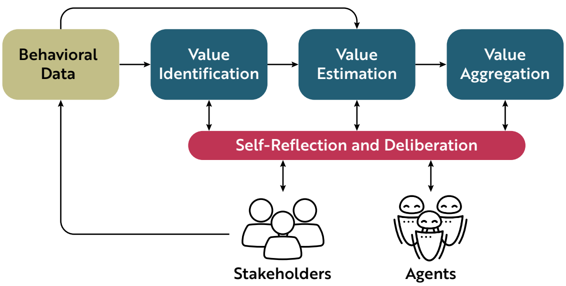

Figure 1.: The hybrid value inference framework.

3. Research Challenges

We identify five interdisciplinary challenges related to hybrid value inference. (1) The value inference process must be *verified* and *validated* [4], i.e., ensuring that it works as intended and to the satisfaction of the stakeholders. Although value inference can be incrementally verified and validated throughout the lifecycle of an STS, it is necessary to define a *satisfaction criterion* for which the results are adequate for being operationalized (e.g., to design policies). (2) Agents must be able to *explain* their actions in an interactive fashion [27], to build trust and guide humans through self-reflection. Further, explanations ought to be actionable [5], with the goal of eliciting appropriate feedback for validating the value inference process. (3) The *resilience* of the process must be measured to guarantee robustness to mistakes [33] and malignant actors [2]. Importantly, given the compositional nature of the proposed framework, resilience should be quantified both for individual steps and for the framework as a whole. (4) Value inference is crucial for sensitive AI applications, e.g., to make life-changing decisions in a healthcare STS. Thus, the *quality of the data* employed in the value inference steps must be curated to guarantee that the process is fair and free of bias [26, 38]. (5) Designing autonomous agents that align with their human users' values is an important step toward trustworthy AI [36, 37]. To this end, the value inference processes must be legitimate [14], providing adequate channels for eliciting stakeholders' consent [37] and dissent [10].

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, 8 2020.

[2] S. Alkobi, D. Sarne, E. Segal-Halevi, and T. Sharbaf. Eliciting Truthful Unverifiable Information. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '18, pages 1850–1852, Stockholm, Sweden, 2018. IFAAMAS.

[3] O. Araque, L. Gatti, and K. Kalimeri. MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowledge-Based Systems*, 191:1–29, 2020.

[4] J. Banks. *Handbook of Simulation*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998.

[5] G. Bansal. Explanatory dialogs: Towards actionable, interactive explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 356–357, New Orleans, LA, USA, 2018. ACM.

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

[7] A. Boixel and R. de Haan. On the Complexity of Finding Justifications for Collective Decisions. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 39–46, Online, 2021. The AAAI Press.

[8] R. L. Boyd, S. R. Wilson, J. W. Pennebaker, M. Kosinski, D. J. Stillwell, and R. Mihalcea. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, ICWSM '15, pages 31–40, Okford, UK, 2015. The AAAI Press.

[9] T. Dietz. Bringing values and deliberation to science communication. *Proceedings of the National Academy of Sciences of the United States of America*, 110(SUPPL. 3):14081–14087, 2013.

[10] R. Dobbe, T. Krendl Gilbert, and Y. Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:1–17, 2021.

[11] 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.

[12] J. González-Pachón and C. Romero. Bentham, marx and rawls ethical principles: In search for a compromise. *Omega*, 62:47–51, 2016.

[13] 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.

[14] S. Grimmelikhuijsen and A. Meijer. Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, 5(3):232–242, 2022.

[15] R. Hadfi and T. Ito. Augmented Democratic Deliberation: Can Conversational Agents Boost Deliberation in Social Media? In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 1794–1798, Online, 2022. IFAAMAS.

[16] M. Hildebrandt. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, 20(1):83–121, 2019.

[17] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, and B. Stein. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '22, pages 4459–4471, Dublin, Ireland, 2022. ACL.

[18] M. Klein. How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium. Technical report, Center for Collective Intelligence, 2012.

[19] 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, pages 1141–1150, New York City, NY, USA, 2009. ACM Press.

[20] R. Lera-Leri, F. Bistaffa, M. Serramia, M. Lopez-Sanchez, and J. Rodriguez-Aguilar. Towards Pluralistic Value Alignment: Aggregating Value Systems through l-Regression. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 780–788, Online, 2022. IFAAMAS.

[21] C. Y. Lim, A. B. Berry, A. L. Hartzler, T. Hirsch, D. S. Carrell, Z. A. Bermet, and J. D. Ralston. Facilitating Self-reflection about Values and Self-care among Individuals with Chronic Conditions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, UK, 2019. ACM.

[22] 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 *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 799–808, Online, 2021. IFAAMAS.

[23] 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 2727–2745, Seattle, WA, USA, 2022. ACL.

[24] 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.

[25] E. Liscio, R. Lera-Leri, F. Bistaffa, R. I. J. Dobbe, C. M. Jonker, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, and P. K. Murukannaiah. Value inference in sociotechnical systems: Blue sky ideas track. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1–7, London, United Kingdom, 2023. IFAAMAS.

[26] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[27] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[28] 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, Auckland, New Zealand, 2020. IFAAMAS.

[29] A. Pommeranz, C. Detweiler, P. Wiggers, and C. M. Jonker. Self-Reflection on Personal Values to Support Value-Sensitive Design. In *Proceedings of the 25th BCS Conference on Human Computer Interaction*, HCI '11, pages 491–496, Newcastle-upon-Tyne, UK, 2011. BCS Learning & Development.

[30] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vinagreiro Alves, E. Vecchione, and S. L. Values and Identities - a policymaker's guide – Executive summary. Technical report, Publications Office of the European Union, 2021.

[31] S. Scheffler. Valuing. In *Equality and Tradition: Questions of Value in Moral and Political Theory*, chapter 7, page 352. Oxford University Press, Oxford, UK, 1st edition, 2012.

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

[33] N. Schwind, E. Demirovic, K. Inoue, and J. M. Lagniez. Partial robustness in team formation: Bridging the gap between robustness and resilience. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 1142–1150, Online, 2021. IFAAMAS.

[34] R. Shortall, A. Itten, M. van der Meer, P. K. Murukannaiah, and C. M. Jonker. Reason against the machine? future directions for mass online deliberation. *Frontiers in Political Science*, 4:1–17, 10 2022.

[35] 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 *HHAI2022: Augmenting Human Intellect*, pages 114–127, Amsterdam, the Netherlands, 2022. IOS Press.

[36] A. M. Singh and M. P. Singh. Wasabi: A conceptual model for trustworthy artificial intelligence. *Computer*, 56(2):20–28, 2023.

[37] M. P. Singh. Consent as a Foundation for Responsible Autonomy. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, AAAI '22, pages 12301–12306, Online, 2022. The AAAI Press.

[38] E. Strickland. Andrew ng, ai minimalist: The machine-learning pioneer says small is the new big. *IEEE Spectrum*, 59(4):22–50, 2022.

[39] Z. Talat, H. Blix, J. Valvoda, M. I. Ganesh, R. Cotterell, and A. Williams. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '22, pages 769–779, Seattle, WA, USA, 2022. ACL.

[40] M. van der Meer, E. Liscio, C. M. Jonker, A. Plaat, P. Vossen, and P. K. Murukannaiah. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect*, HHAI '22, pages 17–31, Amsterdam, the Netherlands, 2022. IOS Press.

[41] S. R. Wilson, Y. Shen, and R. Mihalcea. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *Proceedings of the 10th International Conference on Social Informatics*, SocInfo '18, pages 455–470, St. Petersburg, Russia, 2018. Springer.