# Value Inference in Sociotechnical Systems

Blue Sky Ideas Track

Enrico Liscio
TU Delft, Delft, the Netherlands
e.liscio@tudelft.nl

Roger Lera-Leri
IIIA-CSIC, Barcelona, Spain
rlera@iiia.csic.es

Filippo Bistaffa
IIIA-CSIC, Barcelona, Spain
filippo.bistaffa@iiia.csic.es

Roel I.J. Dobbe
TU Delft, Delft, the Netherlands
r.i.j.dobbe@tudelft.nl

Catholijn M. Jonker
TU Delft/Leiden University
Delft/Leiden, the Netherlands
c.m.jonker@tudelft.nl

Maite Lopez-Sanchez
University of Barcelona
Barcelona, Spain
maite_lopez@ub.edu

Juan A. Rodriguez-Aguilar
IIIA-CSIC, Barcelona, Spain
jar@iiia.csic.es

Pradeep K. Murukannaiah
TU Delft, Delft, the Netherlands
p.k.murukannaiah@tudelft.nl

## ABSTRACT

As artificial agents become increasingly embedded in our society, we must ensure that their behavior aligns with human values. Value alignment entails *value inference*, the process of identifying values and reasoning about how humans prioritize values. We introduce a holistic framework that connects the technical (AI) components necessary for value inference. Subsequently, we discuss how hybrid intelligence—the synergy of human and artificial intelligence—is instrumental to the success of value inference. Finally, we illustrate how value inference both poses significant challenges and provides novel opportunities for multiagent systems research.

## KEYWORDS

Values; Norms; Ethics; Sociotechnical Systems; Hybrid Intelligence

**ACM Reference Format:**

Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. 2023. Value Inference in Sociotechnical Systems: Blue Sky Ideas Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 7 pages.

## 1 INTRODUCTION

Values are the abstract motivations that drive our opinions and actions [81]. The relative importance we ascribe to different values (our *value preferences*) guides our actions. However, how individuals prioritize values is significantly influenced by the socio-cultural environment [23] and the decision context [42, 56]. For instance, consider how the conflict between the values of freedom and safety has shaped the conversation around COVID-19. In sociotechnical systems (STSs) [64, 68], values can be operationalized at both micro and macro [18, 66, 100] levels. At a micro level, an agent ought to align its actions with individuals' value preferences, e.g., by respecting their desire of privacy [1, 63, 65]. At a macro level, values can yield norms to govern the STS [7, 61, 68, 83].

*Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.* 

An important step toward a value-aligned STS is *value inference*, the process of identifying values and reasoning about stakeholders' value preferences. Value inference is a prerequisite for creating systems that align with stakeholders' value preferences. However, inferring values is not trivial. Directly asking humans about their value preferences (e.g., through questionnaires [31, 81]) often leads to incomplete and hypothetical answers that do not reflect real-life behavior [14]. Thus, value preferences ought to be observed from behavior [77]—from our actions and justifications for those.

We identify the fundamental steps of value inference in an STS as (1) *identification* (which values are relevant to a context?), (2) *estimation* (how does each stakeholder prioritize values?), and (3) *aggregation* (what is the societal consensus from individual preferences?).

Value inference cannot be performed solely via computational methods (e.g., machine learning from human behavioral data). Since value reasoning is cognitively challenging [51, 73] and implicit in human thinking [41, 53], values may not be explicitly evident in behavioral data. Further, often humans can express their values only in concrete situations, and values can be emergent [43]. Thus, humans should be systematically guided through the processes of *self-reflection* [53, 72] and *deliberation* [22, 37] to become aware of their value preferences and how they change based on context.

There is an increasing body of AI literature on value inference, focusing on the identification of values [55, 56, 98], the classification of values in text [4, 45, 54], the estimation of individual value preferences [85], and the societal aggregation of value preferences [52]. However, real-world applications often require a combination of these functionalities. In this paper, we offer a holistic view on how the pieces of value inference fit together.

## 2 VALUE INFERENCE

Figure 1 outlines the challenges of value inference as a modular framework consisting of the steps necessary to go from the behavioral data to the individual and aggregated value preferences. The dark blocks in Figure 1 represent *processes* and the light blocks represent the *data* these processes consume or produce.

Our framework's modularization has two advantages. First, the separation of concerns into processes delineates research challenges. Second, the interdependencies between processes expose research

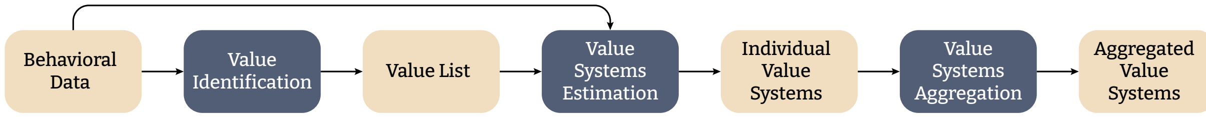


**Figure 1: Value inference processes (dark-colored blocks) and data (light-colored blocks) as a modular framework.**

challenges that can otherwise fall through the gaps. For example, although value identification influences value preferences estimation and aggregation, these connections are largely unexplored.

In our framework, values are inferred from behavioral data. We consider stakeholders' *actions*, e.g., how they choose over competing alternatives [11, 95, 101] or solve a problem [36, 67, 77], as behavioral data. However, value preferences are often implicit in actions and inferring values solely based on actions is difficult. Since language is an important means of value expression, the value preferences underlying our decisions can be observed in the *justifications* we provide for those decisions [31, 80]. Thus, we can exploit the observation of both actions and justifications as the behavioral data that is input to the value inference framework.

***Value Identification.*** Value identification is the process of identifying the set of values relevant to a decision context.

*State-of-the-Art.* Lists of basic human values, applicable across cultures and contexts, have been proposed by ethicists [75, 81] and psychologists [31]. However, such lists are too generic for practical applications [51, 56, 72] and are identified by experts without active stakeholder participation. Value Sensitive Design (VSD) [26] proposes participatory methods for identifying stakeholders' values, e.g., Tuomela et al. [93] employ sensory ethnography to identify the values of users of a smart home energy management system. However, VSD methods usually involve small numbers of stakeholders. Data-driven methods for identifying values have also been proposed, e.g., Wilson et al. [98] (building on Boyd et al. [15]) identify a hierarchy of basic values from user-generate textual content.

*Directions.* Research suggests that not all basic values are relevant to all contexts [51, 56, 72, 81]. Further, an individual's value preferences may not be consistent across contexts [20, 96]. That is, how an individual interprets and prioritizes values depends on *context*. For instance, one might generally value freedom over safety, but prioritize safety over freedom during a global pandemic. Liscio et al. [55, 56] advocate for context-specific values, applicable and defined within a context, arguing that context-specific values are more suitable than basic values for concrete applications (e.g., designing policies). They propose a method for identifying context-specific values, but they involve stakeholders only passively (i.e., by analyzing their deliberation input), while the value identification process is performed by a small group of experts. A comprehensive value list ought to be identified with the active involvement of a representative set of stakeholders and updated over time.

***Value Systems Estimation.*** Values can be ordered according to their subjective importance as guiding principles [81]. Each person has a *value system* that internally defines the importance of values according to their preference and context. In literature, value systems are typically represented as preference rankings over a set of values [83, 85, 102]. Value estimation is the process of determining an individual's value system based on their observed behavior.

*State-of-the-Art.* Value systems are traditionally estimated via surveys [32, 81, 97] over a predefined value list. However, surveys are criticized for not grounding value preferences to a context [56, 72]. VSD methods situate value estimation in a design context by, e.g., showing relevant photos [51, 72] or videos [93]. Yet, the need for human moderation limits the scale in which VSD methods can be applied. In contrast, Inverse Reinforcement Learning (IRL) [67] learns humans' reward functions based on the observed actions, and Cooperative IRL (CIRL) [36] augments IRL with human feedback. However, IRL assumes that humans are aware of their reward functions and is criticized for the infeasibility of estimating an individual's rationality and value preferences simultaneously [60].

*Directions.* As language is our preferred way to express values [31, 80], we envision value systems estimation to be based on both actions and justifications. To this end, Siebert et al. [85] estimate individual value systems from choices and motivations provided in a survey, prioritizing the values expressed in motivations. Recently, several natural language processing (NLP) methods have been proposed for value classification [4, 6, 40, 45, 54], a task identified as necessary for large-scale data processing by the European Commission [79]. With the combination of NLP methods and CIRL, AI agents can estimate value preferences, interactively. However, AI techniques alone are not sufficient, as value preferences are often implicit in our thinking and change over time [41, 60, 91]. Thus, value estimation must be an iterative process, facilitated via self-reflection and deliberation, as we elaborate in Section 3.

***Value Systems Aggregation.*** Value aggregation is the process of aggregating individual value systems into a societal value system, aiming to best represent the societal value canvas [27].

*State-of-the-Art.* The problem of aggregating preferences (e.g., rankings) over a set of objects is widely studied in the computational social choice literature. González-Pachón and Romero [29, 30] show how to aggregate preferences considering an ethical principle that is either utilitarian (i.e., the consensus value system is closest to the majority) or egalitarian (i.e., the consensus value system minimizes the maximum distance with the most displaced individual, hence avoiding the "tyranny of the majority"). To the best of our knowledge, the only work that explicitly addresses value aggregation is by Lera-Leri et al. [52], who propose a method to compute the consensus value system according to any ethical principle, including non-egalitarian and non-utilitarian ones, and test the method on answers to the European Value Survey [95].

*Directions.* Lera-Leri et al. [52] compute one consensus value system according to one ethical principle. However, it is necessary to consider *multiple* consensus value systems when individuals are naturally clustered around different consensuses instead of a single consensus that might not be representative of any individual. From an optimization perspective, this endeavor amounts to solving a *clustering* [25] or, more generally, a *coalition structure generation* problem [17]. Further, value systems can be computed according to *multiple* ethical principles at the same time as individuals might not agree on a single ethical principle for the aggregation problem. Finally, recent research has investigated the importance of providing explanations for decision-making algorithms such as computational social choice [12, 13] and team formation [28]. Explanations are instrumental in collecting stakeholders' feedback, which is critical to validating and improving the value aggregation process.

## 3 HYBRID VALUE INFERENCE

Value inference, as a purely AI task, where a sequence of computational (e.g., machine learning) methods are applied on behavioral data, is not likely to yield good estimates of individual and societal value systems. This is because value preferences are often implicit to humans [41, 53, 91] and are, thus, not easily observable in the behavioral data. Hence, we must actively engage humans, via self-reflection and deliberation, for successful value inference. This makes value inference a hybrid intelligence endeavor [2], requiring human and artificial intelligence to augment each other. Figure 2 shows an overview of the hybrid framework we envision.

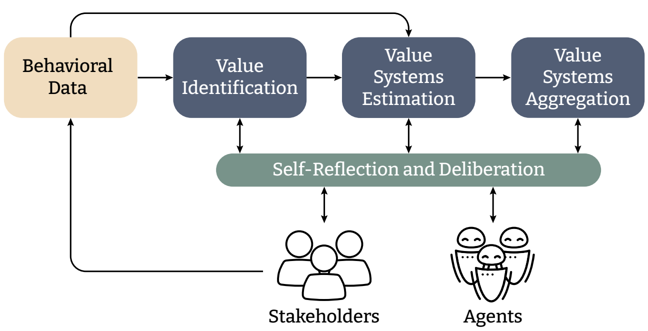


**Figure 2: A hybrid framework, where agents assist humans in self-reflecting and deliberating on value inference processes.**

***Self-Reflection***. Humans must be made aware of values and guided through value reasoning via a process of self-reflection [53, 72]. Self-reflection can be achieved by creating *feedback loops* among the components in our framework. That is, based on the observed behavior and the inferred values, AI agents can interact with humans and help them reflect about their value systems. 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 context, as opposed to asking generic and hypothetical questions over value preferences.

***Deliberation***. In addition to self-reflection, deliberating with others [22, 37] and confronting individuals with different value systems [79] help us in discovering our own value systems. To this end, an increasing number of digital deliberation platforms have been proposed [48, 84]. However, the deliberation quality in unmoderated platforms is often poor, due to polarization and lack of inclusivity [16, 46]. AI-supported human moderation improves deliberation quality [47] but requires large numbers of human moderators. Recently, artificial moderating agents [34, 35] have been proposed to facilitate large-scale deliberation, e.g., a moderating agent can automatically add targeted comments to foster back-and-forth discussions and increase the depth of deliberation.

### A Motivating Example

We introduce an example to demonstrate how self-reflection and deliberation can be fostered in a hybrid value inference framework.

Consider a participatory decision-making situation in which policy makers consult the relevant stakeholders to create COVID-19 regulations. In this case, there is a large variety of stakeholders, including ordinary citizens, healthcare providers, transit authorities, small businesses, and so on. The policy makers seek regulations that satisfy technical constraints (e.g., beds available in the intensive care units) but also align with the stakeholders' value preferences.

To infer the stakeholders' values about potential COVID-19 regulations, policy makers set up a digital deliberation on the issue [38], where participants discuss the impacts of proposed regulations on the healthcare system and the society, and they may vote on different proposals. Artificial agents moderate the discussion by fostering idea sharing and confrontation to increase the deliberation quality.

Value inference can be initially performed based on the participants' behavior on the platform, and subsequently refined through self-reflection and deliberation. Consider that, for a stakeholder, Amber, the estimated individual value system is noticeably different from the aggregated societal value system. Amber's agent investigates this discrepancy. Amber's value system can be incorrectly inferred because (1) the set of identified values does not fully represent Amber's value sentiment (which requires revisiting value identification), (2) Amber's behavior has been misinterpreted (which requires revisiting value estimation), or (3) Amber disagrees that her value system is different from the societal value system (which requires revisiting value aggregation).

Next, Amber's agent can guide her in reflecting on the estimated value systems. For example, if the estimated individual value system is inaccurate because not enough input has been provided in the deliberation, the agent may ask Amber for additional value-laden input through targeted questions (e.g., asking a justification for a specific upvote). The agent can additionally provide explanations about the value inference processes or show the values that were identified from the arguments proposed by Amber. Eventually, the individual value system can remain dissimilar to the aggregated value system. However, through this investigation, Amber is systematically guided by her agent to reflect on her value system.

Finally, Amber and her agent may initiate discussions with other stakeholders and their agents to adjust the value inference processes. For instance, the value list may have to be updated, the NLP model for value estimation may need to account for a minority language, or an aggregation parameter may need to be adjusted to egalitarian instead of utilitarian setting. Importantly, the adjustment of the value inference processes should not be fully automatic. The involvement of relevant stakeholders is essential for meaningful human control [86] on the value inference framework.

## 4 CHALLENGES AND OPPORTUNITIES

We now relate the computational and human-centered research challenges and opportunities associated with hybrid value inference to several emerging research topics in AI. This demonstrates that value inference is a cross-cutting topic that can contribute to and benefit from interdisciplinary research.

***Explainability.*** We identify three connections between explainable AI (XAI) and value inference. First, we emphasize the importance of *interactive* explanations [9, 59, 76], as AI users find a single explanation insufficient [49]. Dialogue-based interactive explanations are a key research challenge for realizing the hybrid value inference framework we envision. Second, explanations are crucial for validating the value inference processes. We envision an AI system that provides explanations for each value inference process with the intent of improving the process itself. To this end, *actionable* explanations (i.e., explanations that humans or agents can act upon) constitute an essential research avenue [9, 44, 74]. Third, we point to the novel challenge of generating *value-based* explanations [99], i.e., natural language clarifications that expose an underlying value reasoning. Such explanations are necessary for communicating the results of value inference to stakeholders.

***Bias, Fairness, and Quality of Data.*** Value inference is crucial for sensitive AI applications, e.g., to make life-changing decisions in a healthcare STS. Therefore, it is essential to guarantee that these decisions do not reflect discriminatory behavior. This amounts to ensuring that the value inference processes are fair and free of bias [50, 57]. This is part of the broader challenge of ensuring the *quality* of the data employed by the value inference processes. To this end, strategies must be devised to *curate* (build, maintain, and evaluate) the datasets involved in value inference. For example, qualitative and quantitative relationships between value datasets can be modeled in a knowledge graph to describe the links between the (context-specific) values inferred in the associated contexts. This is in line with the emerging trend on Data-Centric AI [90], which recommends a focus shift from the models to the underlying data used to train and evaluate models.

***Resilience.*** Value inference is sensitive to misbehavior, as humans may misreport or maliciously misguide their agents when providing feedback. We envision two related research challenges. On the one hand, we can consider how to deter manipulation, which is challenging because it calls for the detection of individual and collective misbehaviors [3]. This would require collaboration with social scientists and economists to design mechanisms for encouraging truthful reporting and feedback that prevent manipulation. On the other hand, we ought to analyze and empirically quantify the *resilience* of the value inference processes when coping with varying populations of misbehaving humans (e.g., by investigating the robustness of the system [62, 82]). For example, the aggregation procedure proposed by Lera-Leri et al. [52] can be sensitive to misreports when computing a consensus according to an egalitarian ethical principle (i.e., with the focus on minorities), since even a single misreport can significantly affect the outcome of the aggregation. Importantly, given the compositional nature of the proposed value inference framework, resilience should be quantified both for individual processes and for the framework as a whole.

***Verification and Validation.*** The results of value inference need to be both verified (i.e., checking whether the processes operate as intended) and validated (i.e., checking whether the system operates to the satisfaction of the users) [8]. Both verification and validation can be performed via hybrid intelligence as described in Section 3. Although value inference can be incrementally verified and validated throughout the lifecycle of an STS, it is necessary to define a moment in which the results are sufficiently satisfactory for being operationalized (e.g., to design policies). Identifying such *satisfaction criterion* is a significant research challenge. This investigation is akin to measuring saturation in qualitative analysis [78], which considers, among other, stakeholders' validation of each value inference process, time and effort required by stakeholders, and evolution of the results (e.g., by identifying asymptotic trends).

***Responsible Autonomy.*** Designing autonomous agents that align with their human users' values is an important step toward trustworthy AI [87, 88]. To this end, the value inference processes must be legitimate [10, 33]. The involvement of stakeholders in the hybrid value inference processes is key to legitimacy, as stakeholders ought to believe that the overall procedure is fair [69]. In particular, consent and dissent are important aspects for ensuring legitimacy [24, 88]. On the one hand, for value inference to be legitimate, the stakeholders must consent to the inference processes. In addition, there must be explicit dissent channels for the stakeholders to question the outcomes of the inference processes. On the other hand, value inference enables a broader understanding of consent, as advocated by Pitkin [70, 71], as not merely seeking a stakeholder's permission but evaluating whether the consented action aligns with the stakeholder's values. Although the concepts of consent and dissent are well-studied in the legal literature [5], computational modelling of these abstractions is an open challenge.

## 5 CONCLUSIONS

Values ought to be considered when building an ethical STS. We explore the challenge of value inference—the endeavor of identifying values and eliciting value preferences both at the individual and societal levels. We outline the components of value inference (identification, estimation, and aggregation), and motivate how a hybrid intelligence approach is instrumental in performing value inference. Finally, we present the related research challenges and opportunities that span multiple AI fields (e.g., MAS and NLP), but also other disciplines including ethics and social sciences.

In practice, value inference is followed by the operationalization of values, both at agent and STS levels, which has been explored in the multiagent community. Values have been used for modeling an individual agent's behavior [1, 63, 65, 94], eliciting appropriate trust [58], plan selection [19], negotiation [7], social simulation [39], and engineering normative systems [61, 62, 83, 89]. We envision value inference and operationalization as actively influencing each other throughout the lifecycle of an STS. An example of such a connection is the evaluation of norm compliance [21, 92], i.e., assessing whether the implemented norms align with the inferred values. Investigating the interdependence of value inference and operationalization is a significant task on its own, which we defer to future work.

## ACKNOWLEDGMENTS

This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No. 952215, and 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. Lera-Leri, Bistaffa, Lopez-Sanchez, Rodriguez-Aguilar are supported by projects VAE TED2021-131295B-C31, funded by MCIN/AEI/10.13 039/501100011033 and NextGenerationEU/ PRTR, VALAWAI (Horizon Europe #101070930), and CI-SUSTAIN (PID2019-104156GB-I00), AUTODEMO (SR21-00329) by Fundación La Caixa, Crowd4SDG (H2020-872944), COREDEM (H2020-785907), Fairtrans (PID2021-124361OB-C33). Maite Lopez-Sanchez belongs to the WAI research group (University of Barcelona) associated unit to CSIC. Lera-Leri, Bistaffa, and Rodriguez-Aguilar are also supported by the "YOMA Operational Research" project funded by the Botnar Foundation.

## 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 (8 2020), 18–28.

[3] Shani Alkobi, David Sarne, Erel Segal-Halevi, and Tomer Sharbaf. 2018. Eliciting Truthful Unverifiable Information. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '18)*. IFAAMAS, Stockholm, Sweden, 1850–1852.

[4] Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL '22)*. ACL, Dublin, Ireland, 8782–8797.

[5] Elizabeth Anderson. 2006. The Epistemology of Democracy. *Episteme: A Journal of Social Epistemology* 3, 1-2 (2006), 8–22.

[6] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowledge-Based Systems* 191 (2020), 1–29.

[7] Reyhan Aydogan, Özgür Kafali, Furkan Arslan, Catholijn M. Jonker, and Munindar P. Singh. 2021. Nova: Value-based Negotiation of Norms. *ACM Transactions on Intelligent Systems and Technology* 12, 4 (2021), 1–29.

[8] Jerry Banks. 1998. *Handbook of Simulation.* John Wiley & Sons, Inc., Hoboken, NJ, USA. 849 pages.

[9] Gagan Bansal. 2018. Explanatory Dialogs: Towards Actionable, Interactive Explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New Orleans, LA, USA, 356–357.

[10] David Beetham. 2001. Political legitimacy. In *The Blackwell Companion to Political Sociology*. Malden and Oxford: Blackwell, New York City, NY, USA, 107–116.

[11] Roland Benabou, Armin Falk, Luca Henkel, and Jean Tirole. 2020. *Eliciting Moral Preferences: Theory and Experiment.* Technical Report. Princeton University. 47 pages.

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

[13] Arthur Boixel and Ulle Endriss. 2020. Automated Justification of Collective Decisions via Constraint Solving. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '20)*. IFAAMAS, Auckland, New Zealand, 168–176.

[14] Dries H. Bostyn, Sybren Sevenhant, and Arne Roets. 2018. Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science* 29, 7 (2018), 1084–1093.

[15] 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 Ninth International AAAI Conference on Web and Social Media (ICWSM '15)*. The AAAI Press, Okford, UK, 31–40.

[16] Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology* 17, 4 (2015), 249–265.

[17] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. 2011. *Computational Aspects of Cooperative Game Theory.* Springer Cham, Cham, Switzerland.

[18] Amit K. Chopra and Munindar P. Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New Orleans, LA, 48–53. https://doi.org/10.1145/3278721.3278740

[19] 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)*. The AAAI Press, Melbourne, Australia, 178–184.

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

[21] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. 2013. No smoking here: Values, norms and culture in multi-agent systems. *Artificial Intelligence and Law* 21, 1 (2013), 79–107.

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

[23] Virginia Dignum. 2017. Responsible Autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. AAAI Press, Melbourne, Australia, 4698–4704.

[24] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. *Artificial Intelligence* 300 (2021), 1–17.

[25] Alessandro Farinelli, Manuele Bicego, Filippo Bistaffa, and Sarvapali D Ramchurn. 2017. A Hierarchical Clustering Approach to Large-Scale Near-Optimal Coalition Formation With Quality Guarantees. *Engineering Applications of Artificial Intelligence* 59 (2017), 170–185.

[26] 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.

[27] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (2020), 411–437.

[28] Athina Georgara, Juan A. Rodriguez-Aguilar, and Carles Sierra. 2022. Building Contrastive Explanations for Multi-Agent Team Formation. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22)*. IFAAMAS, Online, 516–524.

[29] Jacinto González-Pachón and Carlos Romero. 2008. Aggregation of Ordinal and Cardinal Preferences: A Framework Based on Distance Functions. *Journal of Multi-Criteria Decision Analysis* 15, 3-4 (2008), 79–85.

[30] Jacinto González-Pachón and Carlos Romero. 2016. Bentham, Marx and Rawls Ethical Principles: In Search for a Compromise. *Omega* 62 (2016), 47–51.

[31] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*. Vol. 47. Elsevier, Amsterdam, the Netherlands, 55–130.

[32] 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.

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

[34] Rafik Hadfi, Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. 2021. Argumentative Conversational Agents for Online Discussions. *Journal of Systems Science and Systems Engineering* 30, 4 (2021), 450–464.

[35] Rafik Hadfi and Takayuki Ito. 2022. 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)*. IFAAMAS, Online, 1794–1798.

[36] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J. Russell. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS '16)*. Curran Associates, Inc., Barcelona, Spain, 3916–3924.

[37] Catherine Hafer and Dimitri Landa. 2007. Deliberation as self-discovery and institutions for political speech. *Journal of Theoretical Politics* 19, 3 (2007), 329–360.

[38] Johanna Hall, Mark Gaved, and Julia Sargent. 2021. Participatory Research Approaches in Times of Covid-19: A Narrative Literature Review. *International Journal of Qualitative Methods* 20 (2021), 1–15.

[39] Samaneh Heidari, Maarten Jensen, and Frank Dignum. 2020. Simulations with Values. In *Advances in Social Simulation*, Harko Verhagen, Melania Borit, Giangiacomo Bravo, and Nanda Wijermans (Eds.). Springer International Publishing, Cham, 201–215.

[40] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. In

*Proceedings of the Ninth International Conference on Learning Representations (ICLR '21)*. OpenReview.net, Online, 1–29.

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

[42] Patrick L. Hill and Daniel K. Lapsley. 2009. Persons and situations in the moral domain. *Journal of Research in Personality* 43, 2 (2009), 245–246.

[43] Ole Sejer Iversen, Kim Halskov, and Tuck Wah Leong. 2010. Rekindling values in Participatory Design. In *Proceedings of the 11th Biennial Participatory Design Conference*. ACM, Sydney, Australia, 91–100.

[44] Amir Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, New York, NY, USA, 353–362.

[45] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL '22)*. ACL, Dublin, Ireland, 4459–4471.

[46] Hyunwoo Kim, Eun Young Ko, Donghoon Han, Sung Chul Lee, Simon T. Perrault, Jihee Kim, and Juho Kim. 2019. Crowdsourcing perspectives on public policy from stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, Glasgow, UK, 1–6.

[47] Mark Klein. 2012. Enabling Large-Scale Deliberation Using Attention-Mediation Metrics. *Computer Supported Cooperative Work (CSCW)* 21, 4-5 (2012), 449–473.

[48] Mark Klein. 2012. *How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium*. Technical Report. Center for Collective Intelligence. 1–15 pages.

[49] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. In *Proceedings of the Workshop on Human-Centered AI @ NeurIPS (HCAI '22)*. Curran Associates, Inc., Online, 1–23.

[50] Alexander Lam. 2021. Balancing fairness, efficiency and strategy-proofness in voting and facility location problems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '21)*. IFAAMAS, Online, 1806–1807.

[51] 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 City, NY, USA, 1141–1150.

[52] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. 2022. 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)*. IFAAMAS, Online, 780–788.

[53] Catherine Y. Lim, Andrew B.L. Berry, Andrea L. Hartzler, Tad Hirsch, David S. Carrell, Zoë A. Bermet, and James D. Ralston. 2019. 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)*. ACM, Glasgow, UK, 1–12.

[54] Enrico Liscio, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. 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)*. ACL, Seattle, WA, USA, 2727–2745.

[55] 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, 799–808.

[56] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. What values should an agent align with? *Autonomous Agents and Multi-Agent Systems* 36, 23 (2022), 32.

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 1–35.

[58] Siddharth Mehrotra. 2021. Modelling Trust in Human-AI Interaction. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '21)*. IFAAMAS, Online, 1814–1816.

[59] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[60] Sören Mindermann and Stuart Armstrong. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems (NeurIPS '18)*. Curran Associates, Inc., Montreal, Canada, 5598–5609.

[61] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '21)*. IFAAMAS, Online, 907–915.

[62] Javier Morales, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Michael J. Wooldridge, and Wamberto Vasconcelos. 2013. Automated synthesis of normative systems. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '13)*. IFAAMAS, Saint Paul, Minnesota, USA, 483–490.

[63] Francesca Mosca and Jose M Such. 2021. ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '21)*. IFAAMAS, Online, 916–924.

[64] 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)*. IFAAMAS, Auckland, New Zealand, 1706–1710.

[65] 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.

[66] Pradeep K. Murukannaiah and Munindar P. Singh. 2020. From Machine Ethics to Internet Ethics: Broadening the Horizon. *IEEE Internet Computing* 24, 3 (2020), 51–57.

[67] Andrew Ng and Stuart J. Russell. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Cambridge University Press, Stanford, CA, USA, 663–670.

[68] Pablo Noriega, Harko Verhagen, Julian Padget, and Mark D'Inverno. 2021. Ethical Online AI Systems Through Conscientious Design. *IEEE Internet Computing* 25, 6 (2021), 58–64.

[69] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK.

[70] Hanna Pitkin. 1965. Obligation and Consent–I. *The American Political Science Review* 59, 4 (1965), 990–999.

[71] Hanna Pitkin. 1966. Obligation and Consent–II. *The American Political Science Review* 60, 1 (1966), 39–52.

[72] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn M. Jonker. 2011. Self-Reflection on Personal Values to Support Value-Sensitive Design. In *Proceedings of the 25th BCS Conference on Human Computer Interaction (HCI '11)*. BCS Learning & Development, Newcastle-upon-Tyne, UK, 491–496.

[73] 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.

[74] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. ACM, New York, NY, USA, 344–350.

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

[76] Nicholas A Roy, Junkyung Kim, and Neil Rabinowitz. 2022. Explainability Via Causal Self-Talk. In *Advances in Neural Information Processing Systems (NeurIPS '22)*. Curran Associates, Inc., New Orleans, LA, USA, 1–16.

[77] Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin, New York, NY, USA. 352 pages.

[78] 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.

[79] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vinagreiro Alves, E. Vecchione, and Scheunemann L. 2021. *Values and Identities - a policymaker's guide – Executive summary*. Technical Report. Publications Office of the European Union. 12 pages.

[80] Samuel Scheffler. 2012. Valuing. In *Equality and Tradition: Questions of Value in Moral and Political Theory* (1st ed.). Oxford University Press, Oxford, UK, Chapter 7, 352.

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

[82] Nicolas Schwind, Emir Demirovic, Katsumi Inoue, and Jean Marie Lagniez. 2021. 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)*. IFAAMAS, Online, 1142–1150.

[83] Marc Serramia, Maite Lopez-Sanchez, and Juan A. Rodríguez-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.

[84] Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep K. Murukannaiah, and Catholijn M. Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science* 4 (10 2022), 1–17.

[85] Luciano C. Siebert, Enrico Liscio, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon L. Spruit, Jeroen van den Hoven, and Catholijn M. Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*. IOS Press, Amsterdam, the Netherlands, 114–127.

[86] Luciano C. Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van den Hoven, Deborah Forster, and Reginald L. Lagendijk. 2022. Meaningful human control: actionable properties for AI system development. *AI and Ethics* 5, 1 (2022), 1–15.

[87] Amika M. Singh and Munindar P. Singh. 2023. Wasabi: A Conceptual Model for Trustworthy Artificial Intelligence. *Computer* 56, 2 (2023), 20–28.

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

[89] Munindar P. Singh and Pradeep K. Murukannaiah. 2023. Toward an Ethical Framework for Smart Cities and the Internet of Things. *IEEE Internet Computing* X, Y (2023), 1–9. To appear.

[90] Eliza Strickland. 2022. Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big. *IEEE Spectrum* 59, 4 (2022), 22–50.

[91] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. 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)*. ACL, Seattle, WA, USA, 769–779.

[92] Andrea Aler Tubella, Andreas Theodorou, Frank Dignum, and Virginia Dignum. 2019. Governance by glass-box: Implementing transparent moral bounds for AI behaviour. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI '19)*. The AAAI Press, Macao, China, 5787–5793.

[93] Sanna Tuomela, Netta Iivari, and Rauli Svento. 2019. User values of smart home energy management system: sensory ethnography in VSD empirical investigation. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (MUM '19)*. ACM, Pisa, Italy, 1–12.

[94] Sz-Ting Tzeng. 2022. Engineering Normative and Cognitive Agents with Emotions and Values. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22)*. IFAAMAS, Online, 1878–1880.

[95] Tilburg University. 2021. European Value Study. https://europeanvaluesstudy.eu

[96] Caleb Warren, A. Peter Mcgraw, and Leaf Van Boven. 2011. Values and preferences: Defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 2 (2011), 193–205.

[97] K.G. Wilson, E.K. Sandoz, J. Kitchens, and M. Roberts. 2010. The Valued Living Questionnaire: Defining and Measuring Valued Action. *The Psychological Record* 60, 2 (2010), 249–272.

[98] 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.

[99] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. 2021. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artificial Intelligence* 300, 1 (2021), 103554.

[100] Jessica Woodgate and Nirav Ajmeri. 2022. Macro Ethics for Governing Equitable Sociotechnical Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22)*. IFAAMAS, Online, 1824–1828.

[101] WVSA. accessed in 2022. World Value Survey. https://www.worldvaluessurvey.org/wvs.jsp

[102] Luisa M. Zintgraf, Diederik M. Roijers, Sjoerd Linders, Catholijn M. Jonker, and Ann Nowé. 2018. Ordered Preference Elicitation Strategies for Supporting Multi-Objective Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '18)*. IFAAMAS, Stockholm, Sweden, 1477–1485.