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

1Corresponding Author: Enrico Liscio, Delft University of Technology, the Netherlands, e.liscio@tudelft.nl
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]) 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, 41, 40]. 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].

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 in-
teractive 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


