



Value-Aware Multiagent Systems

Nardine Osman^(✉) 

Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Catalonia, Spain
nardine@iiia.csic.es

Abstract. This paper introduces the concept of value awareness in AI, which goes beyond the traditional value-alignment problem. Our definition of value awareness presents us with a concise and simplified roadmap for engineering value-aware AI. The roadmap is structured around three core pillars: (1) learning and representing human values using formal semantics, (2) ensuring the value alignment of both individual agents and multiagent systems, and (3) providing value-based explainability on behaviour. The paper presents a selection of our ongoing work on some of these topics, along with applications to real-life domains.

Keywords: Value awareness · Value alignment · Value learning · Value representation · Real-life applications

1 Value Awareness

There is a pressing need to ensure that the AI systems that we build are not only ethical and beneficial, but also align with our human values. Stuart Russell argues that we should change the goals of the field of AI itself; “instead of pure intelligence, we need to build intelligence that is provably aligned with human values” [19]. This is now known as the value-alignment problem; that is, how to develop systems whose behaviour is aligned with human values.

When studying the alignment of individual agents, the tendency is to reason about the individuals’ decision-making processes [7, 22]. While studying the alignment of multiagent systems has led to reasoning about the norms of that multiagent system [11, 20, 21], as norms have been the traditional means for mediating the behaviour of collectives (organisations, communities, or simple aggregates of individuals).

In this paper, we go beyond the traditional value alignment problem and introduce the notion of value awareness. We define value awareness as follows.

value-aware AI

Noun [U]

/ˈvæl.juː əˈweər ˌeɪˈaɪ/

an AI system that identifies and understands a human’s value system, abides by that value system, and explains its own behaviour and that of others in terms of that value system

This definition lays the roadmap for value engineering research, an emergent field in AI dedicated to the engineering of value-aware systems [14, 15]. First, to **identify and understand** human value systems, the AI should be capable of *learning* relevant values, and *modelling* those values through formal semantics. Of course, value systems may exist on the individual level and the collective level. As such, the AI should also be capable of *aggregating* individual value systems into one that represents the collective. Alternatively, it should be capable of *using agreement mechanisms*, such as argumentation and negotiation, to help the collective agree on their value system.

Second, to **abide** by a value system, *value-alignment* mechanisms are needed for both agents and multiagent systems, such as developing value-aligned decision making and value-aligned norm synthesis. The objective is to ensure behaviour of agents and multiagent systems is aligned with relevant values.

Third, to **explain** one's own behaviour or that of others in terms of value systems, then there is a need for developing *value-based explainability* mechanisms.

We note that value-awareness research is not limited to raising awareness on the agent and multi-agent level, but also raising awareness for humans so they better understand their behaviour. As we see in Sect. 2.3, some of the work in real-life application domains focuses on ensuring humans, like medical professionals or firefighters, better understand which values their (potential) actions are promoting. In other words, the AI also supports humans to make value-informed decisions.

The remainder of this paper provides an overview of a selection of our ongoing work on various topics of this concise and simplified roadmap, along with applications to real-life domains.

2 Selected Contributions

2.1 On Identifying and Understanding Human Values

Value Representation. The first challenge is that of value representation. While work on values in AI is gaining ground, there is no formal model yet for the representation of human values. Furthermore, values in existing AI literature have usually been specified through labels (such as ‘fairness’), without any formal specification of the semantics of those values.

In [12], we propose a value-based taxonomy for the modelling of human values. This allows values to be organised hierarchically, where abstract concepts branch into concrete concepts. Property-based leaf nodes allow us to formally specify value semantics, which enables computational reasoning about values. This is because these property nodes essentially define how a value may be interpreted and assessed. The taxonomy also allows us to explicitly specify value relations and value importance, all of which are crucial elements for deliberating and reasoning over values. Furthermore, we illustrate how the proposed value-based taxonomy is aligned with the values literature in social psychology [13].

Value Learning. One of the main challenges of value awareness engineering is identifying and understanding relevant human values. It is not straightforward for stakeholders to identify and articulate their relevant values. Values are abstract constructs whose exact meaning (semantics) may change over time and context, and even from one person to another. Existing literature illustrates how numerous definitions exist today for values such as fairness or equality. Income inequality alone has been formalised through numerous equations, such as the Gini Index [3], Palma Ratio [4], Theil Index [5], and many many more [8].

We argue that identifying the semantics of values is context dependent, and should involve learning what those values mean for the relevant stakeholders. The learning process should take into account human feedback, which could either be obtained through dedicated user studies or by simply observing the human’s behaviour. Our ongoing work with Hospital del Mar, Barcelona aims at identifying the formulae that best describe the semantics of the four basic bio-ethical values (or principles): beneficence, non-maleficence, autonomy and justice [2]. In this work [17], we are compiling a corpus of patient cases where each case is defined by a set of criteria that specify the patient’s state, an action that is performed on that patient, and the change in the patient’s state after the action is taken. An ongoing user study aims at having doctors annotate these patient cases with information on which of the four values is being promoted, demoted, or not affected. The results of this user study will then be used by an evolutionary strategy algorithm that aims at navigating the space of potential formulae to learn the formulae that best fits the data, and hence, best describes the doctors’ view of how to understand the alignment of each of those values.

2.2 On Abiding by Human Values

Our initial work on the alignment of multiagent systems with human values has mostly focused on the alignment of norms, as norms are what govern the behaviour of these systems. In [21], a basic formal approach for value alignment is presented, where the alignment of a given norm with a given value is assessed by the level of promotion of that value in future states of the world. Norms are understood as changing the future states of the world, and values (such as gender equality) are defined through equations that specify how states of the world may be assessed with respect to that value (such as checking the gender pay gap).

[11] builds on this initial work to present tools for norm synthesis that would optimise for certain values, tools that use the Shapley value concept from game theory to help assess the contribution of the different norms towards promoting those values, as well as tools for checking the compatibility of values under certain norms. The proposed work is tested in a taxing game, where the norms that best promote values like fairness and equality are synthesised and assessed.

[10] proposes an approach that empowers agents by using theory of mind to reason about each others’ values. The proposed mechanism allows agents to analyse norms not only from the perspective of their own value system, but from the perspective of other agents’ value systems. This could eventually help agents when negotiating over the norms that best suit their collective.

Concerning the behaviour of individual agents and their decision making process, we propose in [16] an approach for enhancing automated negotiation mechanisms with social values to help agents reach agreements by considering not only their individual utilities, but also social values such as fairness and equality.

2.3 Real-World Applications

In [1], agent-based simulation is being used to analyse norms (policies) from the perspective of the values of fighting inequality and discrimination. Policies are categorised as aporophobic¹ and non-aporophobic by experts in the field, and agent-based simulation is used to better assess the impact of aporophobia on inequality. Results show that aporophobic policies do in fact lead to larger inequality, compared to non-aporophobic ones.

In the medical field, we are working closely with Hospital del Mar, Barcelona to develop a system that could provide feedback to medical professionals on the alignment of their potential actions with the four basic bioethical values (or principles) [18]. Our proposal is based on analysing the potential outcomes of an action. Alignment is then computed based on the value semantics learnt (see Sect. 2.1). For example, if the value “comfort” was of utmost importance, then actions that lead to states of the world where the patient suffers will not be considered aligned with that value. A multi-objective Markov decision process (MOMDP) is then used to help assess the alignment of entire medical protocols.

A similar approach to our work in the medical field is also being used to help train firefighters. While fire departments have clear and well defined values, it is common to see these values differ from one geographical location to another. Furthermore, new firefighter students usually join with their own value systems, and training over values in agent-based simulations could help them become more aligned with their fire department’s value system.

3 Conclusion

This paper has introduced the notion of value awareness in AI, and presented a concise and simplified roadmap for the development and engineering of value aware AI. It also listed some selected ongoing works covering various research challenges, along with a number of real-life applications.

While the topic of values in AI is becoming more prominent [14, 15], the open challenges are numerous, and the proposed roadmap provides only a glimpse into what future research can delve into. Some of the presented challenges are just starting to get traction, such as the work on aggregating individual value systems into a value system for the collective [9]. It is also evident that research on value-based explainability remains underdeveloped. To address this gap, we intend to

¹ The term aporophobia was coined by the philosopher Adela Cortina to describe having feelings of fear and an attitude of rejection of the poor [6].

build on our symbolic approach for value representation and value-based reasoning, which could provide the foundations for value-based explainability.

Acknowledgments. This work has been supported by the EU-funded VALAWAI (# 101070930) project and the Spanish-funded VAE (# TED2021-131295B-C31) and Rhymas (# PID2020-113594RB-I00) projects.

References

1. Aguilera, A., Montes, N., Curto, G., Sierra, C., Osman, N.: Can poverty be reduced by acting on discrimination? An agent-based model for policy making. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 22–30. AAMAS '24, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2024)
2. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics. Oxford University Press. 8th Edition., New York (2019)
3. Ceriani, L., Verme, P.: The origins of the gini index: extracts from *variabilità e mutabilità* (1912) by corrado gini. *J. Econ. Inequality* **10**, 421–443 (2012)
4. Cobham, A., Sumner, A.: Is it all about the tails? The palma measure of income inequality. Working Paper 343, Center for Global Development (2013). <https://www.cgdev.org/sites/default/files/it-all-about-tails-palma-measure-income-inequality.pdf>
5. Conceição, P., Ferreira, P.: The young person’s guide to the theil index: Suggesting intuitive interpretations and exploring analytical applications. Working Paper 14, University of Texas Inequality Project (UTIP) (2000). <https://utip.gov.utexas.edu/papers/utip.14.pdf>
6. Cortina, A.: Aporophobia: why we reject the poor instead of helping them. Princeton University Press (2022)
7. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: 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, pp. 178–184. *ijcai.org* (2017). <https://doi.org/10.24963/ijcai.2017/26>
8. De Maio, F.G.: Income inequality measures. *J. Epidemiol. Community Health* **61**(10), 849–852 (2007)
9. Lera-Leri, R., Bistaffa, F., Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A.: Towards pluralistic value alignment: Aggregating value systems through l_p -regression. In: Faliszewski, P., Mascardi, V., Pelachaud, C., Taylor, M.E. (eds.) 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022, pp. 780–788. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS) (2022). <https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p780.pdf>
10. Montes, N., Osman, N., Sierra, C.: Perspective-dependent value alignment of norms. In: Osman, N., Steels, L. (eds.) *Value Engineering in Artificial Intelligence*, pp. 46–63. Springer Nature Switzerland, Cham (2024)
11. Montes, N., Sierra, C.: Value-guided synthesis of parametric normative systems. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021, pp. 907–915. ACM (2021)

12. Osman, N., d’Inverno, M.: A computational framework of human values. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 1531–1539. AAMAS ’24, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2024)
13. Osman, N., d’Inverno, M.: Modelling human values for AI reasoning. CoRR abs/2402.06359 (2024). <https://doi.org/10.48550/arXiv.2402.06359>
14. Osman, N., Steels, L. (eds.): Value Engineering in Artificial Intelligence - First International Workshop, VALE 2023, Krakow, Poland, September 30, 2023, Proceedings, Lecture Notes in Computer Science, vol. 14520. Springer (2024). <https://doi.org/10.1007/978-3-031-58202-8>
15. Osman, N., Steels, L. (eds.): Value Engineering in Artificial Intelligence - Second International Workshop Track, VALE 2024, Santiago de Compostela, Spain, October 19, 2024, Proceedings, Lecture Notes in Computer Science, vol. 15356. Springer (In Press)
16. Rodriguez Cima, L., de Jonge, D., Osman, N.: Towards the incorporation of social values in automated negotiation strategies. In: Osman, N., Steels, L. (eds.) Preproceedings of the Value Engineering in AI Workshop track (VALE 2024) at 27th European Conference on Artificial Intelligence (ECAI 2024). Springer Nature Switzerland, Cham (In Press)
17. Rodriguez-Soto, M., et al.: User study design for identifying the semantics of bioethical principles. In: Osman, N., Steels, L. (eds.) Preproceedings of the Value Engineering in AI Workshop track (VALE 2024) at 27th European Conference on Artificial Intelligence (ECAI 2024). Springer Nature Switzerland, Cham (In Press)
18. Rodriguez-Soto, M., et al.: Towards value awareness in the medical field. In: Rocha, A.P., Steels, L., van den Herik, H.J. (eds.) Proceedings of the 16th International Conference on Agents and Artificial Intelligence, ICAART 2024, Volume 3, Rome, Italy, February 24–26, 2024, pp. 1391–1398. SCITEPRESS (2024). <https://doi.org/10.5220/0012588600003636>
19. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin Publishing Group (2019)
20. Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A.: A qualitative approach to composing value-aligned norm systems. In: Seghrouchni, A.E.F., Sukthankar, G., An, B., Yorke-Smith, N. (eds.) Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS ’20, Auckland, New Zealand, May 9–13, 2020, pp. 1233–1241. International Foundation for Autonomous Agents and Multiagent Systems (2020)
21. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perelló, A.: Value alignment: a formal approach. CoRR abs/2110.09240 (2021). <https://arxiv.org/abs/2110.09240>
22. di Tosto, G., Dignum, F.: Simulating social behaviour implementing agents endowed with values and drives. In: Giardini, F., Amblard, F. (eds.) Multi-Agent-Based Simulation XIII - International Workshop, MABS 2012, Valencia, Spain, June 4–8, 2012, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7838, pp. 1–12. Springer (2012). https://doi.org/10.1007/978-3-642-38859-0_1