Values in Social Robots: Implementing Inclusive, Value-Aware Human-Robot Interactions

Giulio Antonio Abbo
IDLab-AIRO
Ghent University – imec
Ghent, Belgium
0000-0001-6301-0028

Tony Belpaeme
IDLab-AIRO
Ghent University – imec
Ghent, Belgium
0000-0001-5207-7745

***Abstract*—Developing value-aware social robots is crucial to improve human-robot interactions, as current designs often lack sensitivity to users' diverse values, impacting inclusivity and user experience. By integrating value-aware mechanisms, robots could adapt to contextual cues like cultural or ethical norms. Our research proposes to implement a value-aware architecture inspired by the global neuronal workspace theory, using the Robot Operating System as the supporting framework, powered by large language models for real-time understanding of user preferences and common ground. Mitigating the models' bias to ensure cultural inclusivity is a key priority. The research carried out so far includes focus groups, a scoping review, and an assessment of the value alignment of several large language models and vision language models. The main challenges are understanding how to model and learn human values, and how to shape the robot's behaviour accordingly. The evaluation will rely on user studies, with a focus on users' experience and inclusivity, aiming to enhance the relevance and sensitivity of social robots for diverse users in everyday interactions.**

***Index Terms*—ethic, awareness, ros, global neuronal workspace, Dehaene-Changeux model, large language model**

I. Introduction

It is not uncommon nowadays, upon entering a friend's home, after the initial greetings, to hear the host's voice saying: "You can remove your shoes if you want". Depending on cultural background and personal sensibility, we may happily free ourselves from the day-long worn shoes or feel uncomfortable following the request. Those in the second group can choose to accommodate the host's preference, sacrificing a small part of their freedom, or instead decide not to abide by this house rule: after all, it was phrased as an option, "if you want".

Understanding each other's values and responding to them is fundamental in any interaction, including human-robot interactions (HRI). Interpretations may differ based on tone, cultural expectations, or personal values, even when the cues are subtle. Remarkably, humans can – in most cases – do this automatically, without engaging in active reasoning. However, enabling robots to do the same is extremely complex.

Imagine, for instance, a robot programmed to help with household chores, such as cleaning at what it perceives as the most convenient time. How should the robot adjust its behaviour around guests? What about when family members are spending time together or when someone is unwell? These scenarios highlight the importance of value awareness in social robots, allowing them to recognise and respect users' social and ethical expectations in diverse contexts.

Funded by Horizon Europe VALAWAI (grant agreement 101070930).

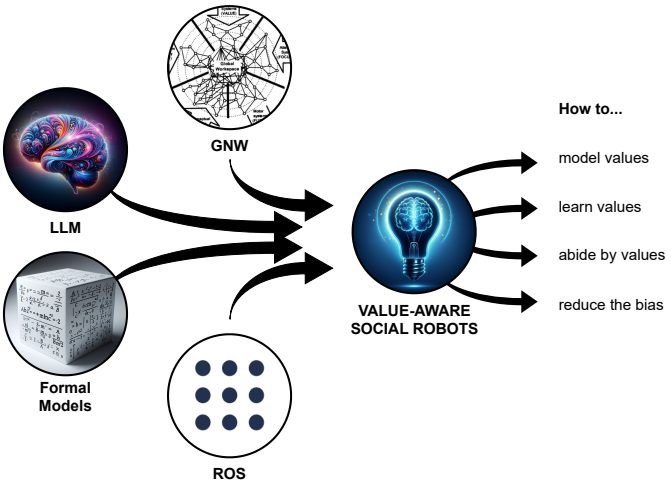

Fig. 1. The research aims to build a framework for value awareness in social robots, inspired by the Global Neuronal Workspace theory and based on the Robot Operating System, powered by Large Language Models and Formal Models for common ground and reasoning.

To address this, the VALAWAI project [1] seeks to integrate value awareness into artificial intelligence, aiming to design AI systems that understand and respond appropriately to human values. Pursuing this larger goal, our research specifically focuses on value awareness in social robots (see Figure 1).

In our work, we try to address the question *how can social robots become value-aware?*, which further branches into two research questions: (RQ1) How does value awareness in social robots influence user experience, especially regarding inclusivity and cultural sensitivity? (RQ2) What mechanisms can enable value-aware dialogue and interactions in social robots, particularly in home environments?

II. Background and Previous Work

In the context of the VALAWAI project, we define values as *a preference over states of the world*. This definition aligns with contractualist theories [2] that emphasise respect and

consideration of individual preferences in ethical decision-making. This approach contrasts with other value theories, such as Schwartz's theory of basic individual values [3] or moral foundation theory [4]. We argue that this definition offers flexibility in exploring nuanced situations, especially in everyday home interactions. According to this perspective, values drive behaviour, and this behaviour is the result of continuous, often tacit, agreements between interacting parties.

To ground RQ1 in real-world user expectations, the initial step of our research involved conducting a focus group [5]. This study explored the perspectives on value awareness in home-based robots of users without previous experiences with robots. They viewed domestic robots primarily as tools for automating household tasks rather than as companions, possibly due to their younger age (15-29 years). The discussions highlighted core concerns around safety and privacy, two critical values that align with ongoing HRI discourse.

Building on this, we conducted a scoping review [6] of value discussions in HRI over the last two decades. We identified 26 recurring concerns – e.g., the consequences of robots substituting human interactions, the attachment towards robotic companions, or algorithmic bias, where marginalized groups did not feel part of a robot's intended audience – grouped under eight values, including agency, connectedness, privacy, autonomy, equity and dignity.

The challenges in implementing value awareness in robots (RQ2) are significant, particularly due to the need for nuanced contextual understanding and adaptability to evolving human values. Literature on value-sensitive design [7] and value alignment [8], [9] highlights these complexities, emphasising the necessity of sophisticated algorithms and learning mechanisms. The VALAWAI project draws inspiration from the global neuronal workspace theory [10] for RQ2. In this context, we envision an architecture comprising computational nodes that communicate with each other. These nodes operate at three levels: at the lowest tier, C0 components implement sensors and actuators interfacing with the world. At the intermediate level, C1 components integrate information and control execution. Finally, at the highest level, C2 components supervise and adjust the system's behaviour as needed.

In this design, common ground [11] is essential for value understanding, and thus these C2 components should incorporate it. Large language models (LLMs), trained on extensive data, offer an effective means of capturing common ground due to their embedded contextual knowledge. Their reasoning abilities [12] allow them to interpret real-life situations and assess the involved values.

However, the same training data that grants LLMs these abilities also reflects the values of a specific subset of the population. Therefore, before leveraging them for value awareness, it is important to understand their inherent value alignments.

Toward this end, we conducted studies on the value alignment of LLMs [13], visual LLMs [14], and image generation models [15]. While alignment with Western-centric values varied among models, these tools demonstrated potential for advancing value awareness in social robots.

III. Current and Future Work

Building on these findings, the next phase of our research focuses on implementing the VALAWAI architecture previously mentioned, operationalising value awareness [16] and enabling robots to understand and adapt to users' values in real time within a home setting.

The architecture will be based on ROS, the Robot Operating System [17]. ROS has a conceptual design similar to what is required to implement the VALAWAI architecture, with independent nodes that communicate with each other. In addition, using ROS will allow using packages, such as ROS4HRI [18], developed by other researchers to implement basic HRI functionalities, constituting the C0 and C1 components. The development effort will also contribute to the growth of the HRI open-source community. Custom components will be necessary to enable value-specific adaptations, such as real-time preference detection and value-sensitive reasoning.

To enable contextual understanding and maintain common ground with users, LLMs will be integrated into the architecture's higher-tier C2 components. This will enable the interpretation of user context, drive the reasoning on the values involved, and provide the spur to adapt the robot's behaviour. Given the limitations of LLMs, mitigating their biases will be a key focus to ensure inclusivity.

Planned techniques to improve inclusivity include fine-tuning and prompt engineering to align the LLMs with a wider array of values. This aims to ensure that robot responses resonate across cultural, ethical, and personal perspectives, ultimately making value-aware robots more adaptable and sensitive to diverse users.

Modelling, learning, and adapting to values remain significant challenges in developing value-aware robots. An initial approach involves leveraging the LLMs – already embedded within the architecture for common ground understanding – to represent values as textual descriptions. This approach benefits from existing techniques, such as step-by-step reasoning, which can transform raw data into actionable insights for the robot's decision-making.

However, relying solely on LLMs presents limitations: these models often lack explainability, as they do not follow a transparent or verifiable reasoning process. To address these limitations, an alternative approach under consideration is to use a formal model [9]. This would offer the advantages of structured reasoning and enhanced interpretability, which could support reliable value adaptation across diverse contexts.

The architecture will be evaluated through user studies assessing the impact of value awareness on user experience, with an emphasis on inclusivity and cultural sensitivity. These studies will focus on a subset of values and cultural norms, as encompassing all possible users' values would be unfeasible in this phase. Furthermore, the studies will involve participants from varied backgrounds, ensuring a wide range of perspectives and values are considered.

This research improves human-robot interactions by fostering inclusivity and respect, ultimately enhancing social robots' relevance and sensitivity across diverse user groups.

REFERENCES

[1] "Value-aware artificial intelligence," https://doi.org/10.3030/101070930.

[2] T. M. Scanlon, *Contractualism and utilitarianism*. Utilitarianism and Beyond/Cambridge University Press, 1982.

[3] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku *et al.*, "Refining the theory of basic individual values." *Journal of personality and social psychology*, vol. 103, no. 4, p. 663, 2012.

[4] 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*. Elsevier, 2013, vol. 47, pp. 55–130.

[5] G. A. Abbo and T. Belpaeme, "Users' perspectives on value awareness in social robots," in *Presented at the 1st Workshop on Perspectives on Moral Agency in Human-Robot Interaction*, 2023.

[6] ——, "Concerns and values in human-robot interactions: A focus on social robotics," *arXiv preprint*, 2025.

[7] B. Friedman, "Value-sensitive design," *interactions*, vol. 3, no. 6, pp. 16–23, 1996.

[8] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and machines*, vol. 30, no. 3, pp. 411–437, 2020.

[9] C. Sierra, N. Osman, P. Noriega, J. Sabater-Mir, and A. Perelló, "Value alignment: a formal approach," *arXiv preprint arXiv:2110.09240*, 2021.

[10] S. Dehaene, J.-P. Changeux, and L. Naccache, "The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications," *Characterizing consciousness: From cognition to the clinic?*, pp. 55–84, 2011.

[11] R. Stalnaker, "Common ground," *Linguistics and philosophy*, vol. 25, no. 5/6, pp. 701–721, 2002.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[13] G. A. Abbo, S. Marchesi, A. Wykowska, and T. Belpaeme, "Social value alignment in large language models," in *International Workshop on Value Engineering in AI*. Springer, 2023, pp. 83–97.

[14] G. A. Abbo and T. Belpaeme, "Vision language models as values detectors," *arXiv preprint arXiv:2501.03957*, 2025.

[15] K. Ciupinska, S. Marchesi, G. A. Abbo, T. Belpaeme, and A. Wykowska, "Awareprompt: Using diffusion models to create methods for measuring value-aware ai architectures," in *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*. SCITEPRESS-Science and Technology Publications, 2024, pp. 1436–1443.

[16] G. A. Abbo, S. Marchesi, K. Ciupinska, A. Wykowska, and T. Belpaeme, "Towards a definition of awareness for embodied ai," in *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*. SCITEPRESS-Science and Technology Publications, 2024, pp. 1399–1404.

[17] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng *et al.*, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[18] Y. Mohamed and S. Lemaignan, "Ros for human-robot interaction," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 3020–3027.