Users' Perspectives on Value Awareness in Social Robots

Giulio Antonio Abbo
giulioantonio.abbo@ugent.be
IDLab-AIRO – Ghent University – imec
Gent, Belgium

Tony Belpaeme
tony.belpaeme@ugent.be
IDLab-AIRO – Ghent University – imec
Gent, Belgium

ABSTRACT

The ability of an artificial interactive system to understand human values and align its actions and decisions with them is important and this is particularly salient in the home environment, thanks to the greater intimacy and more relaxed social structures in private contexts. The technical aspect of building value-aware social robots is further compounded by the fact that people typically see robots as not just a designed artefact, but instead assume intentionality behind a robot's actions. However, little is known about how aware the public is about the need for value-aligned robots, how we will interact with this new and powerful technology and what place we want value-aware robots to take in our lives. This paper explores views on value-aware social robots in the home environment through a series of focus groups. The results show that people are largely unaware and as such unconcerned about a potential mismatch between a robot's behaviour and our own values.

1 INTRODUCTION

Robots have seen a remarkable increase in adoption during the past years, mainly finding applications in the industry where they enable larger throughput, lower costs, and predictable quality [11]. However, their use is not limited to mass production environments and robots are increasingly designed with specific social capabilities that allow more meaningful interactions with people [5]. More recently, for example, social robots have been applied in education [2], providing benefits such as personalised one-on-one teaching. Another relatively new field of application for this technology is medical therapy and care [6], allowing for instance to maintain a high level of engagement with the therapeutic programme [13].

Companion robots are starting to make their way into people's homes, with mixed results [12]. De Graaf et al. suggest that this is mainly due to the negative associations people have with robots [7], which makes it difficult to design a robot that can be widely accepted [8]. Still, AI-powered devices like Amazon Alexa and Google Assistant have been adopted by millions of households, and the expectation is that domestic robots, once the technology is mature enough to meet the exacting expectations of the user, will follow a similar adoption profile.

Yet, social robots and AI-powered devices still lack some necessary features, and one of these is that they are at present completely unaware of human values and the moral consequences of their actions. Indeed, one of the biggest problems concerning AI systems that mimic awareness is that they are not bound by any rules. This often results in calamitous failures, with one well-publicised example being the Tay chatbot which Microsoft had to take offline after less than one day due to it adopting inappropriate language [4, 14].

And while there is considerable effort going towards value alignment in chatbots, with judicially engineered filters preventing the bots from straying too far from acceptable conversation, an awareness of values and social norms would greatly help in the field of social robotics.

Discretion and privacy are values which are particularly relevant for domestic robots. A domestic robot should understand which details about the home and activities in the home should remain secret and which can be shared with others. In another case, a robot assistant could announce its presence before entering a room.

In this paper, we report on a pilot study in which we investigate the views of the "average user" on the topic of value awareness associated with social robots. To this end, a series of focus groups were organised, involving participants of different ages and backgrounds. In particular, the following topics were surveyed:

- Which are the most relevant characteristics of a value-aware agent in the home environment?
- What are the perceived risks and worries associated with a value-aware agent in the home environment?
- What are the attitudes towards the idea of value-aware agents in the home environment?

2 VALUE AWARENESS

The introduction of social robots in the intimacy of the household creates both an opportunity and a need to consider the risks of user manipulation, user deception, breach of privacy, and other behaviours that violate fundamental human or personal values. Making social robots value-aware would ensure that their behaviour remains within moral bounds and enables them to explain why they are taking certain actions or refusing to do them.

This research fits in the context of a large-scale European research project, VALAWAI. The project studies the technology and application of value-aware artificial intelligence in a number of application domains, including social robotics. In this, the AI is able to understand and abide by a value system and explain its own behaviour or understand the behaviour of others in terms of a value system. The project takes the position that value awareness requires a form of artificial consciousness, enabling the AI to become aware of the consequences of its decisions and actions. A number of different interpretations exist for operationalising consciousness, VALAWAI opts to implement the Global Neuronal Workspace model of conscious access by Dehaene and Changeux [9]. The theory is both a high-level neural model offering an explanation for how consciousness emerges from the interaction between different parts of the human brain but also suggests a possible implementation

Cite as: Giulio Antonio Abbo and Tony Belpaeme. 2023. Users' Perspectives on Value Awareness in Social Robots. In *Proceedings of the 1st Workshop on Perspectives on Moral Agency in Human-Robot Interaction*, March 13, 2023, Stockholm, Sweden.

which can be used to create artificial consciousness or, perhaps put more modestly, a form of awareness.

The implementation of the model is however not the subject of this paper. Instead, we wish to form a better understanding of a very unique opportunity provided by building value-aware AI for social robots: the fact that we cannot see social robots as *neutral technology*. When we interact with social robots we take an intentional stance rather than a design stance [15, 18]. We interpret the actions of social robots as resulting from mental states, such as beliefs, desires or intentions. There are a number of poignant observations to be made. One is that taking the intentional stance is largely automatic, it does not require specific priming or expectation setting. The other is that the intentional stance can be amplified by the robot's appearance and behaviour, features that are often part and parcel of social robots as an enhanced intentional stance is often key to the robot's effectiveness. And finally, people take an intentional stance and, by extension, assume the presence of values in the robot even when no such thing has been programmed in.

This creates the need for a better understanding of how users see social robots in domestic settings, in particular how they believe social and domestic robots can align themselves with human values in the context of private homes. In doing so, our work fits in a recent drive to involve the user when designing and developing technology, and specifically when building social robots [e.g. 17, 19].

3 FOCUS GROUP SURVEY

To form an understanding of how naive users see a role for value-aware social robots in domestic contexts, we organised a number of focus groups. Participants were selected based on their potential contribution to our research and specifically on the impact that social robots in the home might have on them [10].

After a preliminary interview, the participants are divided into groups, taking into consideration their age ranges and academic backgrounds. Each group takes part in a separate focus group session. This is done to put participants at ease and make sure that they are not intimidated by speaking in front of someone of a very different age.

The protocol for the sessions is composed of four steps: an ice-breaking activity, the design of a value-aware robot, a critical thinking task, and a review tweet. The details of each phase are reported in the following paragraphs.

At the start of the session, the subjects are invited to sit at a table in a comfortable environment free of distractions. The material for the activities has been previously placed on the table, including a drawing of a house floor plan, sticky notes, paper, and pens. The facilitator introduces and leads the activities, taking notes of the contributions of each participant.

3.1 Ice Breaking Activity

The participants are asked to think of all the tasks that robots and AI-enabled assistants can do in the home. Each participant can write the name of the activity on a sticky note and place it on the relevant area of the floor plan. Participants are asked to explain their propositions as they add them to the map.

To expand the conversation, the facilitator can bring up the topic of people with special needs and how they could benefit from these technologies. For example, the subjects are asked to think of parents with small children, people with motor disabilities, and elderly individuals.

The activity lasts for five minutes. The objective is to get a conversation started and put the participants at ease while talking about a topic that is more familiar than value awareness, while at the same time gathering ideas for the subsequent activities. It is important to note that the facilitator sets no expectations with regard to robots: no definitions nor examples of robots are given, meaning that the responses are only influenced by the participants' earlier notions and conversations during the activity of what robots could do in the home.

3.2 Design a Value-Aware Robot

The facilitator introduces the concept of value awareness, through the example of an intelligent robot that knows whether some information can be disclosed to guests or not, or a vacuuming robot that stops making noise when the owner makes an important call. The subjects are invited to reflect on how one's behaviour changes at someone else's home if they are first-time guests and if they are regular guests.

The participants – possibly divided into small groups – are asked to design a value-aware robot for the home environment. Figure 1 shows the participants engaged in the design phase. They are asked not to worry about know-how and costs associated with their design. This activity is carried out with pens and paper.

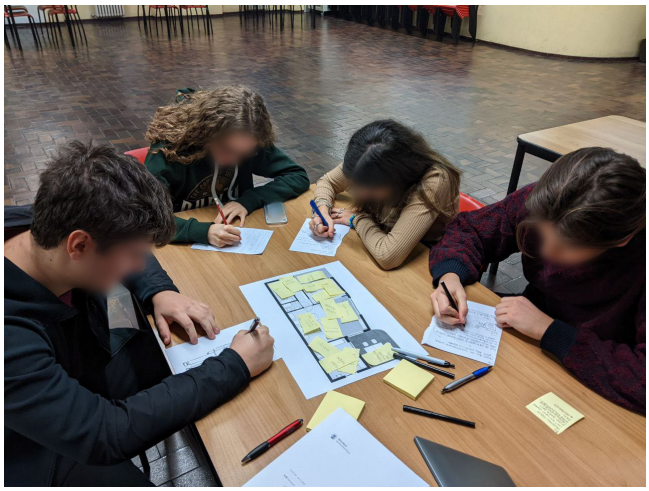

Figure 1: Focus group participants engaged in modelling their value-aware robot

The facilitator takes note of the following aspects: (a) whether the robot is moving or stationary; (b) whether it has an abstract shape or is anthropomorphic; (c) whether it has access to distributed sensors or only to those integrated; (d) how and what it communicates (whether it displays emotions, for example); (e) list of capabilities; (f) whether the robot has to obey someone in particular.

The activity lasts for 10 minutes and each group is then asked to share their design with the others. With this activity, the objective

Table 1: Groups with age statistics

Group	Subjects	Avg	Min	Max	St. Dev.
HS students	5	15.00	15.00	15.00	0.00
Undergraduates	3	20.33	20.00	21.00	0.58
Young w/ CS	5	26.80	25.00	29.00	1.64
Young w/o CS	4	24.75	24.00	26.00	0.96
Total	17	21.71	15.00	29.00	5.07

is to extract which are the most salient characteristics that come to mind when thinking about value-aware social robots.

3.3 Critical Thinking

Next, the groups are invited to reflect on how discretion, privacy, trust, physical safety, emotional considerations, and other factors could play a role in the interaction with the robot they designed, with concrete examples. In addition, they are presented with pictures of existing digital assistant devices, autonomous vacuum cleaners, the Pepper robot, and some examples of interactions with large language models. This allows them to ground their ideation in reality and reduce speculation. After five minutes each group can present their findings to the others. In the end, each group is asked to come up with one moral problem associated with a robot of another group. To complete this activity, the facilitator asks to identify in which rooms of the home the participants would tolerate the presence of a social robot.

The purpose of this activity is to investigate the perceived risks and issues that are associated with the idea of value aware robot in the home environment, in the eye of the average user.

3.4 Review in a Tweet

During the concluding phase, each participant is asked to imagine being a customer that bought the robot just designed and using it daily. They are each given a piece of paper resembling an empty tweet and are tasked to write a short post about something that happened while interacting with the robot, positive or negative. Once the task is over, the participants can share their tweets.

This activity allows probing indirectly the emotions that arise when thinking of possible interactions with a value-aware robot.

4 RESULTS

The focus groups involved 17 volunteers: 10 males and 7 females. The participants took part on a voluntary basis, came from the north of Italy, and had different backgrounds and interests. They were divided into four groups, the characteristics of each group are listed below and reported in Table 1.

- 5 high school students, all 15 years old
- 3 undergraduate students of 20 and 21 years old
- 5 young adults with a computer science background, aged from 25 to 29
- 4 young adults without a computer science background, aged from 24 to 26

4.1 Tasks Associated with AI-Enabled Robots

In total, the ice-breaking activity helped to individuate 21 different tasks associated with robots and AI-powered devices in the home. The most common across sessions were those that had to do with home automation and cleaning. This can be understood in light of the fact that these applications already exist and are now part of the collective culture.

Another recurring topic is that related to a robotic butler, personal trainer, cook, or hairstylist. Indeed, most participants imagined that a robot could take care of the chores in their place, allowing them to have more free time.

In other instances, where it does not completely replace the human in a task, the robot becomes a support device that helps and sustains the user. For example, in the kitchen, an assistant explains recipes, talks and plays music while the person is cooking; while at the entrance it can give useful weather and traffic information when the users are leaving the home, and remind them when they forget to take something with them.

When thinking about people with special needs, screen readers and voice commands came to mind, while smart baby monitors would help adults with young children.

4.2 Appearance and Capabilities of a Value-Aware Robot

Of the 12 designs created, 11 have a physical embodiment, while one is a distributed entity that controls and orchestrates other task-specific robots. The embodied robots are envisioned as barrel-shaped (4/12), with wheels, drawers containing first aid kits and other useful tools, and a tablet located on the top for easy access. Equally popular are the anthropomorphic designs (4/12), they have mechanical features that make them clearly distinguishable from humans. In only one instance the robot looks exactly like a human and cannot be distinguished from one. A quirky design features the robot as a floating bat that can hover out of the way when it is not needed; while another design cannot move and resembles a big cabinet containing all the home appliances.

Only 7/12 robots have access to the home's sensors, the others can rely only on the integrated ones. Voice interaction is ubiquitous and is sometimes complemented by visual interfaces on a tablet, or facial expressions for human-like robots.

In terms of capabilities, 6/12 groups chose cleaning as one of the core functionalities, and 5 wanted a robotic butler capable of attending to the house tasks. Five groups decided to incorporate the functionalities of digital assistants into their design, consequently, their robots can answer general knowledge questions. Companionship emerged as an important aspect for only 4/12 of the proposed designs but was not excluded by the others. Health and security-related features were explicitly mentioned in five robots; in these cases, the robot is a personal trainer that incites the user or is able to perform first aid assistance when necessary.

Concerning values, most of the designs have embedded rules that cannot be broken: among these, the rule to not harm humans, for example. One group envisioned a robot that at first can only carry out basic tasks but quickly learns from the user by imitation. On this note, two groups proposed teaching values to the robots

via voice commands, while one suggested that the robot could also learn and infer values based on data from its sensors.

4.3 Concerns Towards Value-Aware Robots

When asked to reflect on who their robot should obey in the case of conflicting orders, 9/12 groups explicitly mentioned the necessity of an administrator or owner: a figure in charge of and responsible for the robot. Four groups proposed to implement a set of permissions that can be granted to other users, to authorise specific actions. Two groups stated that a value-aware robot should be able to infer who to obey based on the family hierarchy.

To solve situations of conflict, two groups suggested that the robot should freeze and avoid choosing entirely; one group proposed to give priority to less harm, which implies being able to discern and predict which behaviour is going to produce the best results. An interesting approach brought up by one participant was to ask for motivations in the event of conflicting orders. Specifically, in the example, the robot was tasked with cleaning the table by one person, while another requested not to; the robot asked why and the second person answered that he wanted to drink another glass of water: the value-aware robot can then decide who to obey.

During the critical thinking activity, many groups changed their mind about their design. For example, a robot that was supposed to be totally quiet and discreet was changed to ensure that bystanders are aware of its presence. Perhaps predictably, the biggest concern among the participants is that the robot could break the rules and harm someone. In addition, the second most common concern is privacy, with the recurring requirement of not sending data to external services and performing all computations locally.

In general, the participants agreed that there are three main generic concerns with value-aware systems. First, the robot should be aware of what can be said and what must be kept secret. Second, the robot should be aware of what it must not do: for example, it should not hurt humans. Third, it emerged that also how things are said and done plays a fundamental role: a value-aware robot should know how to break the news to someone, and understand when the same action is allowed and when it is frowned upon.

When asked about which rooms should be off-limits for a value-aware robot, the participants took advantage of the *intelligence* of the robot and answered that the robot should be able to understand autonomously or be told which rooms it should not enter based on the user habits. The robot could also behave as a personal assistant would, and knock on the door before entering the room.

4.4 Attitudes Towards Value-Aware Robots

The majority of the short comments gathered during the review-in-a-tweet activity were positive (12/17). The participants imagined themselves returning home and finding that the robot took care of the house and prepared dinner for them (8/17), or having conversations with their robots and being pleased with the outcome (3/17). One tweet was about how the robot protected the home from thieves.

Negative comments were about scepticism towards the robots, their capabilities and safety (3/17). In one instance the robot destroyed an expensive piece of furniture, and in another, the android was found sleeping under the blankets in the owner's bed.

5 CONCLUSION

Our focus group participants predominantly associated domestic robots with the automation of household tasks, while interactivity and companionship did not emerge as an essential application. This is likely due to the younger age of the participants and due to preconceptions set in popular media.

As such, the design of the robot is in most cases utilitarian – shaped like a barrel with wheels, with drawers containing helpful tools – or anthropomorphic but clearly distinguishable from real humans. In all cases, the interaction relies on spoken language, although some designs sport a tablet for manual inputs.

The two major concerns regarding value-aware social robots are related to safety and privacy. The first comes up in scenarios where things go wrong, and the robot *forgets* its values. Most of the concerns voiced by the participants are informed by tropes from science fiction, such as the robot turning into a killer robot. Privacy concerns instead are related to sensor data. While the majority of designs allow the robot to access data from external (home) sensors, participants largely agreed that this data should never be available to third parties. Attitudes towards value-aware social robots are generally positive. For some, however, there is some apprehension about these devices and domestic robots are not automatically trusted. The robot is expected to understand which actions it can take and which topics it can discuss. This choice especially depends on the context, and a value-aware robot is expected to understand the nuances of each situation. Related to this, the robot is expected to understand the implicit relations between people, allowing it to decide, for example, who to obey in case of conflicting orders.

In conclusion, these focus groups revealed that value awareness, while crucial to robotics, is generally not considered by its future users. It falls to researchers to identify the moral problems posed by robots in private contexts, and educate both robot builders and robot consumers [16]. Much like the moral dangers of AI systems and their deployment at large have been widely publicised [e.g. 3], we might need similar initiatives for the moral dangers that robots pose. In the context of domestic and interactive robots, it is key that we do not get stuck in spectacular moral dilemmas –as happened in the self-driving car discussions [1]– but instead approach the subject with nuance befitting the complexities of our home and private environments.

ACKNOWLEDGMENTS

Funded by the Horizon Europe VALAWAI project (grant agreement number 101070930).

REFERENCES

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[2] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social Robots for Education: A Review. *Science Robotics* 3, 21 (Aug. 2018), eaat5954. https://doi.org/10.1126/scirobotics.aat5954

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[4] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: Changing User Needs and Motivations. *Interactions* 25, 5 (Aug. 2018), 38–43. https://doi.org/10.1145/3236669

[5] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social Robotics. In *Springer Handbook of Robotics*, Bruno Siciliano and Oussama Khatib (Eds.). Springer International Publishing, Cham, 1935–1972. https://doi.org/10.1007/978-3-319-32552-1_72

[6] Carlos A. Cifuentes, Maria J. Pinto, Nathalia Céspedes, and Marcela Múnera. 2020. Social Robots in Therapy and Care. *Current Robotics Reports* 1, 3 (Sept. 2020), 59–74. https://doi.org/10.1007/s43154-020-00009-2

[7] Maartje M.A. de Graaf, Somaya Ben Allouch, and Shariff Lutfi. 2016. What Are People's Associations of Domestic Robots?: Comparing Implicit and Explicit Measures. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 1077–1083. https://doi.org/10.1109/ROMAN.2016.7745242

[8] Maartje M. A. de Graaf, Somaya Ben Allouch, and Jan A. G. M. van Dijk. 2019. Why Would I Use This in My Home? A Model of Domestic Social Robot Acceptance. *Human–Computer Interaction* 34, 2 (March 2019), 115–173. https://doi.org/10.1080/07370024.2017.1312406

[9] Stanislas Dehaene, Jean-Pierre Changeux, and Lionel Naccache. 2011. The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications. *Characterizing consciousness: From cognition to the clinic?* (2011), 55–84.

[10] Batya Friedman, David G Hendry, Alan Borning, et al. 2017. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction* 11, 2 (2017), 63–125.

[11] Ruchi Goel and Pooja Gupta. 2020. Robotics and Industry 4.0. In *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*, Anand Nayyar and Akshi Kumar (Eds.). Springer International Publishing, Cham, 157–169. https://doi.org/10.1007/978-3-030-14544-6_9

[12] Anna Henschel, Guy Laban, and Emily S. Cross. 2021. What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports* 2, 1 (March 2021), 9–19. https://doi.org/10.1007/s43154-020-00035-0

[13] Bahar Irfan, Nathalia Céspedes Gomez, Jonathan Casas, Emmanuel Senft, Luisa F. Gutiérrez, Monica Rincon-Roncancio, Marcela Munera, Tony Belpaeme, and Carlos A. Cifuentes. 2020. Using a Personalised Socially Assistive Robot for Cardiac Rehabilitation: A Long-Term Case Study. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 124–130. https://doi.org/10.1109/RO-MAN47096.2020.9223491

[14] Gina Neff and Peter Nagy. 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10 (2016), 4915–4931.

[15] Jairo Perez-Osorio and Agnieszka Wykowska. 2020. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology* 33, 3 (2020), 369–395.

[16] Wendell Wallach and Peter Asaro (Eds.). 2020. . Routledge.

[17] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-based Relational Norms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 353–362.

[18] Eva Wiese, Giorgio Metta, and Agnieszka Wykowska. 2017. Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology* 8 (2017), 1663.

[19] Katie Winkle, Emmanuel Senft, and Séverin Lemaignan. 2021. LEADOR: A method for end-to-end participatory design of autonomous social robots. *Frontiers in Robotics and AI* 8 (2021), 704119.