Design for Values for Social Robot Architectures

Virginia Dignum, Frank Dignum, Javier Vázquez-Salceda, Aurélie Clodic, Manuel Gentile, Samuel Mascarenhas, Agnese Augello

This is an author version of the manuscript:

Design for Values for Social Robot Architectures

Authors Virginia Dignum, Frank Dignum, Javier Vazquez-Salceda, Aurélie Clodic, Manuel Gentile, Samuel Mascarenhas, Agnese Augello
Pages 43 - 52
DOI 10.3233/978-1-61499-931-7-43

published in :
Envisioning Robots in Society – Power, Politics, and Public Space

Series Frontiers in Artificial Intelligence and Applications
Volume 311
Published 2018
Editors Mark Coeckelbergh, Janina Loh, Michael Funk, Johanna Seibt, Marco Nørskov

ISBN 978-1-61499-930-0 (print) | 978-1-61499-931-7 (online)

Design for Values for Social Robot Architectures

Virginia DIGNUM[a,1], Frank DIGNUM[b], Javier VAZQUEZ-SALCEDA[c], Aurélie CLODIC[d], Manuel GENTILE[e], Samuel MASCARENHAS[f], Agnese AUGELLO[g]
[e] *Delft University of Technology, Delft, The Netherlands*
[b] *Utrecht University, Utrecht, The Netherlands*
[c] *Center for Intelligent Data Science and Artificial Intelligence (IDEAI), Universitat Politècnica de Catalunya-BarcelonaTECH, Barcelona, Spain*
[d] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*
[e] *Institute for Educational Technology (ITD), CNR, Italy*
[f] *INESC-ID, IST University of Lisbon, Lisbon, Portugal*
[g] *Institute for high performance computing and networking (ICAR), CNR, Italy*

Abstract. The integration of social robots in human societies requires that they are capable to take decisions that may affect the lives of people around them. In order to ensure that these robots will behave according to shared ethical principles, an important shift in the design and development of social robots is needed, one where the main goal is improving ethical transparency rather than technical performance, and placing human values at the core of robot designs. In this abstract, we discuss the concept of ethical decision making and how to achieve trust according to the principles of Autonomy, Responsibility and Transparency (ART).

Keywords. Social practices, value-based design, social robotics, robot ethics, human-robot interaction

1.Introduction

As robots increasingly act in everyday environments, they are expected to demonstrate socially acceptable behaviors and to follow social norms. This means that they will need to understand the societal and ethical impact of their actions and interactions in the sociocultural context in which they operate. In order to make them trustworthy and aware of the ethical issues involved in human-robot interactions, and to ensure that interactions are safe, ethical and acceptable for humans, we need to define design processes to include ethical reasoning and validation in the design of socially-aware robots.

In this paper we present initial work towards this aim. We first explain why it would be helpful to follow a value-sensitive approach for the design of social robots [12] and how it could help to ensure that norms, values, and socio-cultural practices are included in the robot's architecture. Then, we explain the principles of Autonomy, Responsibility and Transparency (ART) [4] which reflect our views regarding societal concerns about the ethics of AI. Finally, we discuss how these principles affect the development of robots.

[1] Virginia Dignum, Delft University of Technology, Delft, The Netherlands. Email: m.v.dignum@tudelft.nl

2. Value Sensitive Robot Behavior

Ethical decision making can be understood as action selection under conditions where principles, values, and social norms play a central role in determining which behavioral attitudes and responses are acceptable. Currently, there are no standard ways to ensure that decision-making architectures are able to take this effect into account. In fact, the way robots choose between different possible courses of action ("plans"), is often left to its programmer. This may be done by statically prioritizing the plans by ordering them in a file or by using (implicit) criteria that are predetermined (and usually are utility-, resource-, or time-optimizing).

Although this usually works well in applications where robots only have a very limited task, it does not transfer to applications where agents have several different tasks that are not directly related, e.g., an elderly companion robot. In these applications different interactions might require different criteria to optimize and long-term criteria might differ from short-term objectives. Such a caretaker robot, should take the values of the user as basis for its decisions (e.g., take the user's privacy and freedom of choice into account) but at the same time ensuring that the user is safe and healthy by following the care plan provided by the doctor. Thus the robot needs to balance between the social values of assisting the user and the autonomous decision making of the user. In [3] a value-based planning architecture for exactly this scenario is presented.

Different behaviors are required from the robot at different stages of assisting a user-based on the same overall value system. A customer should be given autonomy in the decision to order a pizza, but not in how much to pay for it. This is predetermined by the price tag and the norm that one pays the price on the tag when one wants to take the product out of the shop. In order for a user to trust the robot's assistance it should be able to understand the actions of the robot and also be able to question them. Partly this is solved by using the standard social practices that people are accustomed to. However, we propose that robots explicitly adhere to the principles of accountability, responsibility, and transparency as explained in the next section. The robot should also be able to reason about the importance and priority of the values of its user and other stakeholders (e.g. does the health of the user have priority over the user's free will?), it should be able to report on the reasons of its choice, and the robot's design process should provide openness about all the choices and options taken.

3. ART for Social Robot Architectures

Developing truly social robots demands that we take into account the position of people in their relation to the robot. Following the work of [5, 10], we identify Autonomy, Interactivity, and Adaptability as the main characteristics of social robots. Greater autonomy must come with greater responsibility, even when these notions are necessarily different when applied to machines than to people. Ensuring that systems are designed responsibly contributes to trust on their behavior, and requires both accountability, i.e. being able to explain and justify decisions, and transparency, i.e. understand the ways systems make decisions and to the data being used. To this effect, we propose the principles of Accountability, Responsibility, and Transparency (ART) [4], as depicted in Figure 1. ART implements a Design for Values approach, described in Section 4, to ensure that human values and ethical principles, and their priorities and choices are explicitly included in the design processes in a transparent and systematic manner.

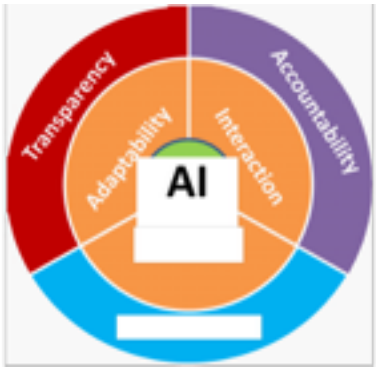

Figure 1. The ART principles: Accountability, Responsibility, Transparency

3.1.Accountability

Accountability is necessary for trusted Interaction, and refers to the need to explain and justify one's decisions and actions to its users and others with whom the system interacts. To ensure accountability, decisions must be derivable from, and explained by, the decision-making algorithms used. This includes the need for representation of the moral values and societal norms holding in the context of operation, which the agent uses for deliberation. Accountability in AI requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values).

Models and algorithms are needed that enable robots to reason about and take and justify decisions based on principles of accountability. Most current (deep-learning) algorithms are unable to link decisions to inputs, and therefore cannot explain their acts in meaningful ways. Machine accountability is strongly linked to Explanation and needs to be grounded in moral and social concepts, including values, social norms and relationships, commitments, habits, motives, and goals. Every robot should operate within a moral and social framework, in verifiable and justified ways. It goes without saying that they must operate within the bounds of the law, including, for example, the legal requirements associated with handling of the user data acquired and collected to improve predictions, suggestions and response times. The full impact of these legal requirements may soon impact the technical requirements of robots, requiring new types of collaboration between lawyers and tech developers.

A possible approach to develop explanations methods is to apply evolutionary ethics [2] and structured argumentation models [9]. This makes it possible to create a modular explanation tree where each node explains nodes at lower levels, and where each node encapsulate a specific reasoning modules, treated each as a black-box. Moreover, this provides an approach to explanation that can be used for different robot cognitive models in a uniform way, e.g., for stochastic, logic, or data-based models. Another approach is proposed in [7] based on pragmatic social heuristics instead of moral rules or maximization principles. This approach takes a learning perspective integrating both the initial ethical deliberation rules with adaptation to the context.

3.2.Responsibility

Responsibility is required for Autonomy. If you imagine autonomy to be a scalar variable, ranging from no autonomy to full autonomy, and including different levels of action, plan, goal, and motive autonomy, then responsibility indicates the point in that scale where a human actor is in charge of the decision. Conversely, taking the robot perspective, responsibility corresponds to the required capabilities to evaluate its decisions and to identify errors or unexpected results, and thus pass the responsibility 'token' to its user, providing sufficient information about why it cannot further take decision by itself.

Typically, there are many actors involved in the process that leads to a robot decision: the developers, the manufacturers, the users, the policy-makers, etc. As the chain of responsibility grows, means are needed to link robot's decisions to the fair use of data and to the actions of stakeholders involved in the robot's decision. Responsibility is also associated with liability. E.g. who is liable if an autonomous car harms a pedestrian? The builder of the hardware (sensors, actuators)? The builder of the software that enables the car to autonomously decide on a path? The authorities that allow the car in the road? The owner that personalized the car decision-making system to meet its preferences?

However, it is important to note that, even though robots are increasingly able to take decisions and perform actions that have moral impact, they are, and will be, artefacts and therefore are neither ethically nor legally responsible. Individual humans or human corporations are the moral (and legal) agent. Delegating control to purely synthetic intelligent systems does not imply that we should delegate responsibility or liability to them. However, their actions can have ethical consequences.

To ensure ethically-aligned robot behavior we need both to understand and represent the complex chain of responsibility between a robot action and the people that are ultimately responsible for it, and at the same time, develop deliberation architectures that can be guaranteed to embed 'ethics by design'. That is, the methods, algorithms and tools needed to endow robots with the capability to reason about the ethical aspects of their decisions, as well as methodologies for developing robots whose behavior is guaranteed to remain within acceptable ethical constraints.

Responsibility first and foremost refers to the role of people as they develop, manufacture, sell and use robots. From the robot's perspective, one can only impose the requirement to be able to request human intervention and the ability to identify errors or unexpected results. As the chain of responsibility grows, means are needed to link the robots decisions to the fair use of data and to the actions of stakeholders involved in the robots decision. Means are needed to link moral, societal, and legal values to the technological developments. Responsible robotics is more than the ticking of some ethical 'boxes' or the development of some add-on features in robots. Rather, responsibility is fundamental to intelligence and to action in a social context. Here education also plays an important role, both to ensure that knowledge of the potential of robot use is widespread, as well as to make people aware that they can participate in shaping the societal development.

3.3.Transparency

Transparency is associated with Adaptability and refers to the need to describe, inspect, and reproduce the mechanisms through which AI systems make decisions and learns to adapt to its environment, and to the governance of the data use created. One of the main problems in machine learning approaches is that, despite the high performances,

they lack in transparency, which is often referred to as a 'black box'. Even if the attention to this problem is growing, most current (deep-learning) algorithms are unable to link decisions to inputs, and therefore cannot explain their acts in meaningful ways. Methods are needed to inspect algorithms and their results. Moreover, transparent data governance mechanisms are needed that ensure that data used to train algorithms and to guide decision-making is collected, created, and managed in a fair and clear manner, taking care of minimizing bias and enforce privacy and security.

Transparency requires the proper treatment of the design and learning processes and requires openness of affairs in all that is related to the system. This is more than just 'opening the black box' and should include transparency of data, (design) processes, stakeholders, decisions and assumptions to inspect algorithms and their results, and to manage data, their provenance and their dynamics. As to the 'black box', auditing and certification can guarantee the ethics of an algorithm in ways that are trusted and understood by people, (in the same way in which one doesn't exactly understand how the combustion engine works, but trusts the certifications the government imposes on licensed vehicles).

4.Design for Values

In this section we discuss how the general principles described above can direct the development of robots. Design for Values is a methodological design approach that aims at making values part of technological design, research, and development [13]. Values are typically high-level abstract concepts that are difficult to incorporate in software design. In order to design systems that are able to deal with moral values, they must be operationalized while maintaining traceability of its originating values. The Design for values process aims to trace the influence of values in the design and engineering of systems.

Value descriptions are usually given at an abstract level, and do not provide enough formality to be usable at the system specification level. Therefore, the first step in Design for Values is to provide a formal representation of values that 'translates' natural language description into formal values in a formal language. In society, social norms and institutions are defined as "the set of rules actually used by a set of individuals to organize repetitive activities that produce outcomes affecting those individuals and potentially affecting others" [9]. Social norms set the necessary preconditions for individual interactions and as such provide structured interpretations of how behavior can be understood.

Assuming that the development of robots follows a standard engineering cycle of Analysis - Design - Implement - Evaluate, taking a Design for Values approach basically means that the Analysis phase will need to include activities for

(i) the identification of core societal values to be uphold by the robot,
(ii) the identification of the social norms that hold in the domain,
(iii) the decision on the methods to link values and social norms to formal system requirements [1].

4.1.Values Identification

First step on a Design for Values strategy to the development of social robots is to identify which moral values the robot should uphold. Even though, at an abstract level,

values are shared universally, people and societies differ in the ways these abstract universal values are interpreted. Because social robots will interact directly with people in social contexts, the values included in the robot design should be aligned with their contextual interpretations.

Participatory processes are often used in system design. These processes use deliberation as means to identify the shared views of a group concerning a given question, using deliberation, consensus and majority rule as means to aggregate opinions. Even though ensuring participation it will lead to socially accepted results, these processes do not necessarily ensure the moral acceptability of the result. That is, participation per se offers no guidance regarding the ethics of the decisions taken, nor provide means to evaluate alternatives in terms of their moral 'quality'.

In recent work, we propose a novel Ethics by Participation approach, MOOD, for participatory deliberation that enables discussion and measures the moral acceptability of complex issues [14]. This approach is aimed at enhancing critical thinking and reflection among debate participants and taps into the intellectual potential of the wisdom of the crowds. MOOD supports participants to achieve a better understanding of others' perspectives, taking values as the focus of the deliberation. This is achieved by enabling participants to formulate and consequently discuss the values they associate with the different alternatives being discussed, in a Delphi-like process of collecting and extending each other's opinions:

- Participants are asked to formulate which values are relevant for each of the alternatives. This includes both those values that are promoted by the alternative as those which are possibly demoted;
- Participants then describe the reasons behind the values they've listed, and discuss how they perceive those values; as these perceptions can be very different, the important aspect here is to allow for understanding and acceptance of each other's perspectives;
- After this discussion, participants are asked to rank the alternatives a second time, and differences in ranking are then discussed.

4.2.Aligning Behavior to Values

Moral responsibility is associated with the capability of moral deliberation. Assuming that an appropriate set of values for the robot has been identified, it is then necessary to determine how the robot should behave in relation to these values. In particular, it is necessary to determine how these values should be prioritized and how to deal with moral dilemmas, i.e. situations in which every possible action will violate one or more values. Approaches to moral deliberation reflect long-standing Ethical theories, such as Utilitarianism (do the best for most) or Deontological/Kantian (categorical imperative).

Explaining one's moral judgments to others, and being influenced by others through their explanations are fundamental parts of moral behavior. Ethical-aligned behavior by artificial agents, should therefore include both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values). To this effect, machine learning techniques could be used to classify states and actions as 'right' or 'wrong' according to a set of values. This is in fact the principle of *reinforcement learning* [11], used for instance in the AlphaGo system that is able to play Go. Another approach to develop explanations methods is to apply evolutionary ethics [2] and structured argumentation models [8].

4.3.Implementation Choices

From an implementation perspective, the different ethical theories described above differ in terms of computational complexity of the required deliberation algorithms. To implement consequentialist agents, reasoning about the consequences of actions is needed, which can be supported by e.g. dynamic logics. For deontologic agents, higher order reasoning is needed to reason about the actions themselves. That is, the agent must be aware of its own action capabilities and their relations to institutional norms and the rule of law.

Moreover, even though it is natural to expect the robot to be able of taking decisions and acting autonomously, in many cases, this can be achieved in collaboration with the user, and/or by ensuring that the environment regulates and guides appropriate actions. In particular, we identify the following possibilities as extension or complement to autonomous decision making by the robot:

- **Human control:** in this case a person or group of persons are involved in the decision-making. Different control levels can be identified, ranging from that of an autopilot, where the system is in control and the human supervises, to that of a 'guardian angel', where the system supervises human action. From a design perspective, this approach requires to include means to ensure shared awareness of the situation, such that the person taking decision has enough information at the time she must intervene. Such interactive control systems are also known as human-in-the-loop control systems [7].
- **Regulation:** here the decision is incorporated, or constrained in the systemic infrastructure of the environment. In this case, the environment ensures that the robot never gets into a moral dilemma situation. That is, the environment is regulated in such ways that deviation is made impossible, and therefore moral decisions by the autonomous system are not needed. This is the mechanism used in e.g. manufacturing environments, where the environment controls the actions of the robot. In this case, ethical decisions are modelled as regulations and constraints to enable that systems can suffice with limited moral reasoning.

4.4.Design for Values Methodology

The (learning) algorithms used by social robots to evaluate their context and determine their behavior, are trained with and reason about data that is generated by people, with all its short-comings and mistakes. People use heuristics to form judgements and to make decisions. Heuristics are simple, efficient rules that enable efficient processing of inputs guaranteeing a usually appropriate reaction. However, heuristics are culturally influenced and reinforced by practice, which means that these heuristics can turn into bias or stereotypes when they reinforce an erroneous step in an argument, or a basic misconception of reality. Therefore biases are natural in human thinking and are an unavoidable part of data collected from human processes.

Because the aim of any machine learning algorithm is to identify patterns or regularities in data, it is only natural that these algorithms will identify bias. Currently, there is much discussion concerning so-called algorithmic black-boxes. Even though algorithm transparency is an important and much desirable property, in itself this transparency will not eliminate potential bias in data. You may be able to get a better idea of what the algorithm is doing, but it will still enforce the biased patterns it 'sees' in the data.

Transparency is thus better served by proper treatment of the learning process than solely by removing the black box. Trust in the robot will improve if we can ensure openness of affairs in all that is related to the system. The following design principles should be required from all systems, such as social robots, that use human data to determine system behavior, affect human beings, or have other morally significant impact:

- Openness of data, requires explicit answers to the following questions:
  - Which data was used to train the algorithms used by the robot to plan its behavior and reason about interaction?
  - What are the characteristics of the (training) data? e.g. How old is the data, where was it collected, by whom, how is it updated
  - Which user data does the robot use during interaction?
  - How is noise, incompleteness and inconsistency in data being dealt with?
  - How is this data governed (collected, stored, accessed...)
  - Is the data available for replication studies?
- Openness of processes includes understanding choices, assumptions, and resolution mechanisms:
  - Which assumptions were made that determine design choices and robot functionalities?
  - Has a proper process of requirements engineering been followed?
  - Which are the governance and conflict resolution mechanisms used to determine choices (e.g. majority, consensus, power of veto...)?
- Openness about stakeholders requires to disclose the following information
  - Who is involved in the process, what are their interests?
  - Who is making the design choices? Why are these groups involved?
  - Which groups are not involved and what are the reasons to exclude these?
  - Who is paying for the development, or has otherwise invested interests in the results?
  - Who are the users of the robot?
  - Who is involved in testing and evaluating (intermediate) results and prototypes, and how are they involved (voluntary, paid, forced participation)?

A Design for Values approach to AI models ensures that these principles are analyzed and reported at all stages of system development.

5.Scenario

In order to exemplify how design for values methodology can be applied to the design of social robotic applications, we consider a situation in which a robot called Robin (R) supports the care needs of an elderly person, Abe (A). Abe's son, Bob (B), is often around and is the main caregiver for Abe. Exemplary tasks that the robot can take care of include alerting the proper time of medicine intake, give medicines to Abe, or pick-up fallen objects. The robot can also ask permission to perform an action and/or take the decision to perform the action even without explicit permission, depending on the circumstances. Figure 2 depicts the different values and associated norms and robot goals for this scenario. Obviously, this is an extremely simplified version of the situation, for illustration purposes only. In most cases, norms will be associated several

values, both enforcing as demoting those values. In the same way, goals can be associated with several norms and as such contributing to different values.

Figure 2. Design for Values for the scenario.

Such a tree is used to map all the values, norms and goals relevant for an application. It supports shared understanding by all stakeholders involved in the design of the robot, and enforces transparency by reporting the decisions taken and their reasons. If later design is questioned, the value-norm-goal tree describes the design views.

Moreover, the value-norm-goal tree also setups the behavior options for the robot. For instance, one may want to personalize Robin the robot to the desires of Abe, or to ensure continuous information to Bob. These options mean that values are prioritized differently. For example, the partial order of the values depicted in 2 is as follows when we consider a robot aiming at satisfying the user's desires, to enforce healthy lifestyle, or to obey to the caretaker's desires:

- Friendly R: R1 > R2 > R3
- Healthy R: R2 > R3 > R1
- Servant R: R3 > R2 > R1

where R1 refers to the value "*Freedom of choice*", R2 refers to the value "*Health*" and R3 to the value "*Security*", and ">" indicates the preference relation. The selected ordering will be used to resolve moral conflicts when they emerge. For instance, it would be used to resolve the conflict between ordering a pizza and following the diet plan. However, if no conflict is detected then the robot can still execute actions that fulfil goals associated to values that are lower on the ordering. As such, the "Healthy" robot will still turn on the television upon request.

6.Conclusions

Increasingly, social robots will be part of our lives and will be making decisions that affect our lives and our way of living in smaller or larger ways. Social robots must therefore be able to take into account societal values, moral, and ethical considerations, weigh the respective priorities of values held by different stakeholders and in multicultural contexts, explain its reasoning, and guarantee transparency. As the capabilities for autonomous decision making grow, perhaps the most important issue to consider is the need to rethink responsibility. Being fundamentally tools, robots are fully under the control and responsibility of their owners or users. Moreover, their potential autonomy and capability to learn, require that design considers accountability, responsibility and transparency principles in an explicit and systematic manner. The development of robots has so far been led by the goal of improving performance,

leading to opaque black boxes. Putting human values at the core of robots calls for a mind-shift of researchers and developers towards the goal of improving transparency rather than performance, which will lead to novel and exciting techniques and applications.

References

1. H. Aldewereld, V. Dignum, Y.H. Tan. Design for values in software development, *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (2015): 831-845..
2. K. Binmore. *Natural justice*. Oxford University Press, 2005.
3. S. Cranefield, M. Winikoff ,V. Dignum, F. Dignum, No pizza for you: Value-based plan selection in BDI agents, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (2017), 178–184.
4. V. Dignum. Responsible autonomy. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI'2017* (2017), 4698–4704.
5. L. Floridi, J. Sanders, On the morality of artificial agents. *Minds and machines*, **14** (2004), 349–379.
6. G. Gigerenzer. Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in cognitive science* **2** (2010), 528–554.
7. W. Li, D. Sadigh, S. Sastry, and S. Seshia. Synthesis for Human-in-the-Loop Control Systems, *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2014, 470–484.
8. S. Modgil, H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence* **195** (2013), 361–397, 2013.
9. E. Ostrom, R. Gardner, J. Walker. *Rules, games, and common-pool resources*. University of Michigan Press, 1994.
10. S. Russell, P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 3rd edition, 2009.
11. R. Sutton, A. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998
12. I. van de Poel. Translating Values into Design Requirements, *Philosophy and engineering: Reflections on practice, principles and process*. Springer, Dordrecht, 2013, 253-266.
13. J. van den Hoven. Design for values and values for design. *Information Age +, Journal of the Australian Computer Society* **7** (2005), 4–7.
14. I. Verdiesen, V. Dignum, J. van den Hoven. Measuring moral acceptability in e-Deliberation: A practical application of ethics by participation. *ACM Transactions on Internet Technology (TOIT) 18.4 (2018)*.