Educational Requirements for Positive Social Robotics

Johanna Seibt

Abstract Social robotics does not create tools but social ‘others’ that act in the physical and *symbolic* space of human social interactions. In order to guide the profound disruptive potential of this technology, social robotics must be repositioned—we must reconceive it as an emerging interdisciplinary area where expertise on social reality, as physical, practical, and symbolic space, is constitutively included. I present here the guiding principles for such a repositioning, “Integrative Social Robotics,” and argue that the path to culturally sustainable (value-preserving) or positive (value-enhancing) applications of social robotics goes via a *redirection* of the humanities and social sciences. Rather than creating new educations by disemboweling the humanities and social sciences, students need to acquire full disciplinary competence in these disciplines, as well as the new skill to direct these qualifications toward membership in multidisciplinary developer teams.

So-called social robots are artificial agents designed to move and act in the physical and symbolic space of human social interactions—as automated guides, receptionists, waiters, companions, tutors, domestic assistants, etc. According to current projections, already by 2025 there will be a US\$100 billion market for service robots, and by 2050 we might have automated 50% of all work activities (McKinsey 2017). As economists gladly usher in the “automation age” (ibid.), it is crucial to be clear on a decisive difference between digitalization and unembodied AI on the one hand and embodied social AI on the other. The ‘as if’ of simulated sociality in embodied AIs (social robots) captivates us with disconcerting ease. For the first time in human cultural history, we produce, for economic reasons, *technological artifacts that are no longer tools* for us—we are building “social others”.

More than a decade of human-robot interaction (HRI) research reveals how willingly humans engage with social robots, practically but also at the affective

J. Seibt (✉)
Aarhus University, Aarhus, Denmark
e-mail: filseibt@cas.au.dk

level, and these research results raise far-reaching theoretical and ethical questions. Should the goings-on between humans and robots really *count as* social actions? Will we come to prefer the new “friends” we *made* to human friendships we need to cultivate? If robots display emotions, which increases the fluidity of social interactions (Fischer 2019), will we be able to learn *not* to respond with moral emotions (sympathy)? Or should robots have rights? Will social robots de-skill us for interacting authentically with other people?

Decisions pertaining to the use of social robots are not only highly complex—as the example of sex robots may illustrate most strikingly—but also bound to have momentous socio-cultural repercussions. However, research-based policy making on social robots is currently bogged down in a “triple gridlock of description, evaluation, and regulation” (Seibt et al. 2020a, b) combining descriptive and prescriptive uncertainty. Currently, we do not know precisely how to describe human reactions to robots in non-metaphorical ways; the lack of precise and joint terminology hampers the comparability of empirical studies; and the resulting predictive uncertainties of our evaluations make it impossible to provide sufficiently clear and general regulatory recommendations.

Robotics engineers increasingly appreciate that their creations require decisional competences far beyond their scientific educations. Single publications (see, e.g., Nourbakhsh 2013; Torras 2018) and the recent IEEE “Global Initiative on Ethics of Automated and Intelligent Systems” document impressive efforts “to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity” (IEEE n.d.).

While we should wholeheartedly endorse these efforts, the question arises whether new educational standards, requiring obligatory ethics modules in engineering educations, will be sufficient. More precisely, these efforts may not suffice as long as we retain the current model of the research, design, and development process (RD&D model) in social robotics.

According to the current RD&D model, roboticists, supported by *some* relevant expertise from other disciplines, create an *object* (robot) which is supposed to *function* across application contexts. What these objects *mean* in a specific application context hardly comes into view. Even if engineering students were to acquire greater sensitivity for the importance of ethical considerations—e.g., as a first step, the insight that “ethical considerations” go beyond research ethics (data handling, consent forms, etc.)—it is questionable that even a full year of study (using the European nomenclature: a 60 ECTS module) could adequately impart competences for responsible decision-making about the symbolic space of human social interactions. *The symbolic space of human interactions is arguably the most complex domain of reality we know of*—structured not only by physical and institutional conditions but also by individual and socio-cultural practices of “meaning making,” with dynamic variations at very different time scales. Even a full master’s study (4–5 years) in the social sciences or the humanities is barely enough to equip students with professional expertise (analytical methods and descriptive categories) necessary to understand small regions or certain aspects of human social reality.

In short, given that the analysis of ethical and socio-cultural implications of social robotics applications requires *professional expertise* in the social sciences or humanities, which short ethics modules cannot provide, our current RD&D model for the development of social robotics applications places responsibilities on the leading engineers that they cannot discharge.

The way forward is thus to modify the RD&D model for social robotics applications. In line with design strategies such as “value-sensitive design” (Friedman et al. 2002), “design for values” (Van den Hoven 2005), “mutual shaping” (Šabanović 2010), and “care-centered value-sensitive design” (Van Wynsberghe 2016), the approach of “Integrative Social Robotics” (ISR) (Seibt et al. 2020a, b) proposes a new developmental paradigm or RD&D model that is tailor-made for our current situation. As a targeted response to the triple-gridlock and the socio-cultural risks of social robotics, ISR postulates an RD&D process that complies with the following five principles (for details, see *ibid.*):

(P1) The *Process Principle*: The product of a RD&D process in social robotics are not objects (robots) but social interactions.

This principle makes explicit that social robotics generates (not instruments but) new sorts of interactions that (1) are best understood as involving forms of asymmetric sociality and, even more importantly, (2) belong into a complex network of human social interactions. This shift in our understanding of the research focus of social robotics immediately motivates the following principle:

(P2) The *Quality Principle*: The RD&D process must involve, from the very beginning and throughout the entire process, expertise of all disciplines that are directly relevant for the description and evaluation of the social interaction(s) involved in the envisaged application.

The Quality Principle demands the constitutive involvement (a) of expertise pertaining to the functional success of the envisaged application—e.g., health science, gerontopsychology, education science, or nursing—but also (b) of expertise from researchers from the social sciences and humanities, in order to ensure adequate analyses of the socio-cultural interaction context of the application. Robotics engineers currently develop applications often guided by their own ethical imagination and social competence alone, which is, in our current situation, a highly problematic underestimation of the tasks and risks involved. One central condition for a deeper understanding of the socio-cultural interaction context is formulated in the following principle:

(P3) The *Principle of Ontological Complexity*: Any social interaction I is a composite of (at least) three components $\langle I_1, I_2, I_3 \rangle$, which are the agent-relative realizations of interaction conceptions as viewed from the perspective of (at least) two interacting agents and an external observer. The interaction conceptions of each of the directly interacting agents (i.e., I_1 and I_2) consist in turn of three perspectival descriptions (from a first-, second-, and third-person view) of the agent’s contribution, while the interaction conception of the external observer is only from a third-person perspective. The RD&D process in social robotics must envisage, discuss, and develop social interactions on the basis of a perspectival account of social interactions.

A differentiated understanding of social interactions along these lines allows us to analyze in detail how people experience their interactions with a robot; in-depth investigations into these partly pre-conscious processes of “sociomorphing” and their associated conscious phenomenology (Seibt et al. 2020b) are crucial elements in the continuous evaluation of the robot’s (physical, kinematic, and functional) design, as demanded by the following principle:

(P4) The *Context Principle*: The identity of any social interaction is relative to its (spatial, temporal, institutional, etc.) context. The RD&D process must thus be conducted with continuous and comprehensive short-term regulatory feedback loops (participatory design) so that the new social interaction is integrated with all relevant contextual factors. Importantly, participatory feedback continues for quite some time after the “placement of the robot,” until a new normality has formed.

Since according to ISR Principle 2, the Quality Principle, researchers from the humanities and, in particular, ethicists are included in the RD&D process, the Context Principle ensures that both individual preferences of the stakeholders are taken into account as well as the interests of society at large. The Context Principle acknowledges the complexity of social reality, but also expresses a commitment to a combined empirical (bottom-up) and normative (top-down) determination of “what matters” in the given application context. This is reinforced by the following principle:

(P5) The *Values-First Principle*: Target applications of social robotics must comply with a specification of the *Non-Replacement Maxim*: *social robots may only do what humans should but cannot do*. (More precisely: robots may only afford social interactions that humans should do, relative to value V, but cannot do, relative to constraint C.) The contextual specification of the Non-Replacement Maxim is established by joint deliberation of all stakeholders. Axiological analyses and evaluations are repeated throughout all stages of the RD&D process.

The formulation of the *Values-First Principle*, as well as of the Non-Replacement Maxim, reflect a commitment to an account of values in line with classical pragmatism, values are realized in interactions and experiences, in a strictly context-dependent fashion. Accordingly, the ISR approach begins with an extended intake phase with careful field research on the value landscape of the interaction context before inserting the new interaction and is accompanied by an ongoing value dialogue of all stakeholders (including the ethicists among the researchers/developers) that reaches into the phase of the “new normal” many months after the placement of the robot.

The Value-First Principle of ISR can be applied with two different levels of ambition. First, applications can be selected that comply with the Non-Replacement Maxim relative to a given context C and raise a value V1 that within that context C ranks higher than a value V2 which is negatively affected by the application. For example, a companion robot for elderly residents of a nursing home which assists in establishing telecommunication with the resident’s family will raise the autonomy of the resident; given that the nursing home cannot offer continuous human assistance and an in-person visitation plan with the family is in place, the resident may rank the increase in her autonomy higher than the loss of some human contact with staff.

Such an application, which warrants an increase of a value relative to the axiological ranking within the context, can be considered as *culturally sustainable*.

Second, working with a more restrictive selection filter, developer teams working with ISR might choose only to pursue applications that (1) comply with the Non-Replacement Maxim relative to a *class* of contexts and (2) raise a value V1 without affecting the axiological ranking in this context class. For example, delivery of conflict mediation via a genderless humanoid telecommunication robot can apparently increase the likelihood that conflict parties find constructive resolutions for gender-charged conflicts (Druckman et al. 2020); this seems to be due to the fact that the genderless robot does not provide stimuli for gender-related perceptual bias, a feature that human mediators should but (typically) cannot exhibit (Skewes et al. 2019). Such an application, where a human-robot interaction is *intrinsically valuable*, can be considered an instance of *positive social robotics* (in extension and further specification of the notion of “positive computing” (Calvo and Peters 2014)).

I have described the five principles of the ISR approach in some detail here in order to motivate the following three claims about what culturally sustainable and positive social robotics would imply for future higher education:

Claim 1: Reposition social robotics: Given the currently incalculable socio-cultural risks of widespread use of social robots, we need new RD&D models (such as ISR) where (a) the expertise of the humanities and social sciences is centrally and constitutively included from beginning to end, and (b) applications are developed in a value-driven fashion.

Claim 2: No new educations, refurbish and redirect the humanities: We need to conceive of social robotics as an emerging interdisciplinary area of research—but we are not yet in the position to introduce new bachelor’s or master’s educations in this area. Culturally sustainable or even positive social robotics currently still needs the combination of *professional* expertise in all relevant disciplines, i.e., expertise that is acquired in the course of full professional educations in these disciplines. However, if humanities educations in the future aim to qualify students for work in social robotics developer teams, the humanities also need to change their self-image from purely reflective to (also) pro-actively engaged areas of knowledge and adjust curricula accordingly.

Claim 3: Create interfaces: In order to equip students with suitable skills for work in interdisciplinary developer teams for culturally sustainable or positive social robotics, we need to establish *interface modules*. The core disciplines of the emerging interdisciplinary area of social robotics, such as robotics, anthropology, philosophy, (geronto)psychology, design, health, nursing, and education, should include modules of the course work of one semester (typically three courses) where students acquire skills in interdisciplinary communication and collaboration, as well as some basic introductions to terminology and methods of other core disciplines, with focus on ethics.

Claim 1, postulating the constitutive inclusion of the humanities and social sciences, is a direct consequence of the second principle of ISR, the “Quality Principle”: it is scientifically irresponsible to build high-risk applications without involving those disciplines that can professionally assess the risks of value corruptions (the corruption of public discourse and fact-finding practices by social media algorithms illustrates the consequences of irresponsible technology development).

Claim 3 follows from the fifth principle of ISR, the “Values-First Principle,” which offers a gradual exit, piece-meal, from the current triple-gridlock of research-based regulation. If developer teams concentrate on positive, value-enhancing

applications, we can gradually learn more about human encounters with social robots while reducing the potential risks—building the applications “we could want anyway.” However, as my commentaries on the Values-First Principle may convey, value discourse and the analysis of values require a *certain mindset, a tolerance for ambiguity, complexity, and deep contextuality*, that cannot be learned by top-down rule application or as “how-to” lateral transfer. While an education in the humanities cultivates the development of such special mindsets, this is not so in other disciplines. A preparatory course in applied ethics with well-chosen examples (see, for instance, course material in connection with Torras 2018) can gradually train students from non-humanities disciplines to *understand* more about epistemic attitudes and standards of reasoning in complex normative domains, but it will likely not suffice to acquire them. Vice versa, as currently explored in a supplementary educational module on “Humanistic Technology Development” at Aarhus University, in order to create epistemological and communicational interfaces from the other side, humanities students should learn a bit of programming and design and build a rudimentary robot.

Finally, let us consider claim 2, the claim that we should better *not* push for new educations, at least not now. This claim might be surprising, in view of interdisciplines like bioinformatics or nanotechnology where the identification of a new research field led quickly to the introduction of new educations. On the other hand, as Climate Research and Systems Biology illustrate, there are interdisciplinary areas the complexity of which, relative to our current understanding, requires full disciplinary competences. If we begin to interfere with a domain as intricate as social reality, and care for more than money, we cannot afford cheap solutions with cross-disciplinary educations stitched together cut-and-paste. Given the complexity and contextuality of social reality, given the demand for value-driven applications based on quality research (see ISR Principles 2–5), social robotics needs developer teams with full expertise rather than mosaic knowledge in order to create futures worth living.

References

- Calvo, R.A., Peters, D. (2014). Positive computing: technology for wellbeing and human potential. MIT Press.
- Druckman, D., Adrian, L., Damholdt, M.F., Filzmoser, M., Koszegi, S.T., Seibt, J., Vestergaard, C. (2020). Who is Best at Mediating a Social Conflict? Comparing Robots, Screens and Humans. *Group Decis. Negot.* <https://doi.org/10.1007/s10726-020-09716-9>
- Fischer, K. (2019). Why Collaborative Robots Must Be Social (and even Emotional) Actors. *Techné Res. Philos. Technol.* 23, 270–289.
- Friedman, B., Kahn, P., Borning, A. (2002). Value sensitive design: Theory and methods. University of Washington technical report 02–12.
- IEEE, n.d. IEEE SA – The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [WWW Document]. URL <https://standards.ieee.org/industry-connections/ec/autonomous-systems.html> (accessed 10.28.20).

- McKinsey Global Institute, A Future that Works, Automation, Employment and Productivity, 2017, <https://www.mckinsey.com/mgi/overview/2017-in-review/automation-and-the-future-of-work/a-future-that-works-automation-employment-and-productivity>
- Nourbakhsh, I.R. (2013). Robot futures. MIT Press.
- Šabanović, S. (2010). Robots in society, society in robots. *Int. J. Soc. Robot.* 2, 439–450.
- Seibt, J., Damholdt, M.F., Vestergaard, C. (2020a). Integrative social robotics, value-driven design, and transdisciplinarity. *Interact. Stud.* 21, 111–144.
- Seibt, Johanna, Vestergaard, C., Damholdt, M.F. (2020b). Sociomorphing, Not Anthropomorphizing: Towards a Typology of Experienced Sociality, in: *Culturally Sustainable Social Robotics--Proceedings of Robophilosophy 2020, Frontiers of Artificial Intelligence and Its Applications*. IOS Press, Amsterdam, pp. 51–67.
- Skewes, J., Amodio, D.M., Seibt, J. (2019). Social robotics and the modulation of social perception and bias. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20180037.
- Torras, C. (2018). *The Vestigial Heart*. MIT Press.
- Van den Hoven, J. (2005). Design for values and values for design. *Information age* 4, 4–7.
- Van Wynsberghe, A. (2016). Service robots, care ethics, and design. *Ethics Inf Technol* 18, 311–321.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.