Designing Socially Assistive Robots

A Relational Approach

Tony J. Prescott[1][0000-0003-4927-5390], and Julie M. Robillard[2][0000-0001-5765-4927]

[1] Department of Computer Science, University of Sheffield, UK
[2] Division of Neurology, Department of Medicine, University of British Colombia, Canada

t.j.prescott@sheffield.ac.uk

Abstract. There is increasing interest in social robots as assistive technologies to support a wide range of potential user groups. Nevertheless, the widespread use of robots has been challenged in terms of their efficacy and benefits as well as the ethics of employing robots in social roles. For instance, it has been suggested that robots are incapable of being truly social and therefore that any use of social robots as assistive technology is intrinsically deceptive. This contribution addresses this controversy, building on a relational view of human-robot interaction, which asserts that sociality has less to do with the essential natures of the human and robot actors involved, and more to do with the patterns and consequences of their interaction. From this starting position we consider and explore four design principles for social robots and compare/contrast these with the view of design "transparency" that robots should behave to reveal their true machine nature.

Keywords: Social Robots, Robot Ethics, Deception, Relational Ethics.

1 Introduction

Social robots are increasingly used for assistive applications across the lifespan with populations across the entire spectrum of vulnerability [1-3]. Examples include as distraction devices for children undergoing painful procedures, as communication aids in children with autism, as mental health interventions in adults and children, and as interventions to reduce agitation in older adults living with dementia. Until recently, social robots were seen as somewhat futuristic and largely existed in the realm of research. However, the Covid-19 pandemic has substantially accelerated their use in a wide range of contexts from education to healthcare as part of a drive to maintain social connectedness while limiting close physical contact. As such, questions surrounding the ethics of social robots and the nature and morality of human-robot relationships are more pressing than ever, with important implications for how social robots are designed and employed in real-world settings. Here we consider different approaches to human-robot relationships, describe the key components of a relational approach, and propose four evidence-based design principles for the ethical design of social robots.

2 The Relational Approach in Robot Ethics

The relational view in robot ethics argues for a move away from essentialist (or substantialist) notions of what is a human, what is a robot, and what it means for them to have a relationship. Instead the relational view proposes that what matters are the patterns and consequences of social interactions between humans and robots [4-6], including their meaning and significance to the people involved, and their wider impact on social and relationship contexts [3]. This view can be seen as an alternative to more essentialist conceptions that seek to define what is (and what is not) a human (or a robot) in terms of fundamental character or attributes irrespective of context (see [7]). Essentialist views can be attractive ways to frame and explore certain ethical questions as they chime with many of our intuitions (for instance that all humans share a common "nature") and language (which emphasizes objects acting on each other as opposed to systems with multiple interacting elements), however, they can be criticized on metaphysical grounds [8], for supporting outdated ideas of the human that can be exclusionary [9], and for failing to recognize the changing nature of our humanity, including through our interactions with our technologies [10, 11]. The relational view in technology ethics, on the other hand, is part of a broader interactivist turn in the social, cognitive and information sciences (e.g. [8, 12, 13]) that sees the units involved in a social transaction (e.g. humans and robots) as deriving ''their meaning, significance, and identity from the (changing) functional roles they play within that transaction" ([8] p. 287).

While the debate between relational and essentialist views continues, we consider that it is useful to explore and set out some of the implications of the relational view for the design of assistive technologies, particularly, those such as robots, and other social AIs (e.g. smart speakers), that purport to have some social function, and whose benefits are considered to arise, at least in part, through their sociality.

The possibility that a robot could be deemed to be social is hotly contested. For instance, Sparrow [14] has argued than robots (and similar devices) are incapable of sociality, and that to present them as otherwise is intrinsically deceptive and morally deplorable. Reflecting on similar views, has led some authors to propose that, to be ethical, robots should be designed such that their machine nature is transparent. To enable this transparency, it is suggested that the user should be reminded, occasionally, if not continuously, that the device is a machine controlled by algorithms rather than a "genuine" social actor [15, 16].

Central to this debate is the question of what it means to be deceptive. We follow Danaher [17] who defines deception as involving "the use of signals or representations to convey a misleading or false impression" (p. 118). In robotics, deception is most often held to be about portraying a misleading impression about qualities that humans have, and that robots do not (or in principle could not) have. We might summarize these as anthropomorphic qualities, or more specifically, a sub-class of anthropomorphic qualities that are deemed controversial, most often psychological phenomena such as emotions, intentions, and self-awareness (in contrast, physical features such as having a head, two arms and two legs, are rarely considered deceptive or problematic). If robots exhibit qualities or functionalities that are viewed as deceptive, the further question is whether this is, indeed, unethical. Broadly speaking, we see three general positions set out in table 1. The first two are broadly similar only differing in what they see as the solution to the ethical "problem" of social robotics. We identify with the third of these positions (of which there are multiple versions), which begins from a more nuanced view on the nature of deception in robotics.

Table 1. Views on deception and ethics in social robotics.

Are social robots deceptive?	Is this unethical?	What should we do about it?	Example authors
Yes	Yes	Avoid building or using them altogether	Sparrow [14]; Turkle [18]
Yes	Yes	Design it out, or minimize it through transparency	Boden et al. [19]; Wortham & Theodorou [16]
Not necessarily	Depends on the nature of the deception	Design to avoid damaging forms of deception	Shim & Arkin [20]; Sorell & Draper [21]; Danaher [6, 17, 22]; Prescott & Robillard [3, 7, 23]

Determining whether social robots are deceptive by nature requires reflection on our understanding of sociality. To rule out the possibility that an artefact could ever be social seems exclusionary given that we do not yet have a clear understanding of human sociality or how it is generated [7]. Moreover, embodied cognitive science is forcing a rethink about the nature of sociality as something that arises not in individuals but in the interactions that occur between them [12]. Applied to robots, this suggests that they need not have self-understanding, or intrinsic social competencies or properties to be authentically social [24].

Nevertheless, we might agree that present-day robots are not social in the same way that people are. If so, is it possible to defend the deliberate creation of an impression of human-like sociality (as, for example, artificial personal assistants strive to do)? A key idea here is that the tendency to anthropomorphize objects and devices occurs widely and pre-dates robotics and artificial intelligence [25, 26]. For example, we anthropomorphize dolls, cars, even trees and mountains.

A related point is that we may be able to distinguish different forms of deception, and that some of these may not be unethical. For example, anthropomorphism, has been described as being "honest" where it exploits people's tendency to view artefacts as social actors, and does so overtly and for their benefit, using anthropomorphic features to provide a more engaging or effective interaction (for example, to provide navigation instructions in a vehicle, or to promote the effectiveness of a therapy) [27]. However, anthropomorphism can be seen as "dishonest" where it is used to deliberately misdirect attention or conceal a robot capability. For example, to pretend that the robot is unable to see a person because its artificial eyes appear closed even while continuing to observe them with a covert camera [27, 28].

Danaher [17] has argued that some forms of honest anthropomorphism are not unethical even though they may be deceptive. Analyzing different forms of deception employed by robots, Danaher describes an "ethical behaviorist" approach, according to which judgements about whether a robot's anthropomorphic behavior is permissible should be based on superficial observables—including the robot's appearance, utterances and actions—and not on any presumptions about the presence or absence of human-equivalent robot inner states. This is termed "superficial state deception". As Danaher puts it:

"According to ethical behaviorism, if a robot appears to have certain capacity (or intention or emotion) as a result of its superficial behavior and appearances, then you are warranted (possibly mandated) in believing that this capacity is genuine. In other words, if a robot appears to love you, or care for you, or have certain intentions towards you, you ought, ceteris paribus, to respond as if this is genuinely the case. [...] simulated feeling can be genuine feeling, not fake or dishonest feeling. Consequently, if ethical behaviorism is true, then superficial state deception is not, properly speaking, a form of deception at all." (p. 122-3).

Danaher's position can be likened to a strong version of the relational perspective (e.g. [24]), that is, that what manners is that the robot's behavior, over the duration of its interactions, is consistent with its social utterances and expressions. This is a stronger constraint than you might at first imagine as explored further below.

3 Design Principles for Social Robots

Based on the above, and from a relational standpoint, we believe it should be possible to define design principles for ethical social robots. As an initial effort, we propose the following:

1. Promote **contextual integrity**: This principle advocates co-design of robot social capabilities for the role that the robot will fulfil and alignment of the robot's behavior and capabilities with expectations and norms. Nissenbaum [29] introduced the notion of "contextual integrity" in the context of a framework for the design of sociotechnical systems, applying it particularly to concerns around information privacy; however, the idea has broad generality. Its application to robotics has been discussed further by Kaminski et al [27]. The key idea is that the capabilities and behavior of a robot should be judged in terms of their appropriateness to the context in which it is used. For example, if we encounter a social robot that is waiting tables in a restaurant, we might reasonably expect that it would enter the room unannounced, observe where people are sitting and approach them safely, monitor ongoing conversation and diner behavior for an appropriate point at which to intercede with and offer of service and so-on. The same robot, but in a home setting, might be required to observe quite different social etiquette, for example, never entering certain rooms,

asking before entering others, not using cameras or microphones at certain times of the day, or in some situations, unless specifically directed to do so.

2. Develop **honest anthropomorphism**: This principle requires that we evaluate the benefits and risks of anthropomorphic features and make decisions on their permissibility accordingly. “Superficial state deception” can be acceptable if consistent with expectations and norms; “hidden state deception”, such as where the robot conceals a covert feature that might violate contextual integrity, is unacceptable. Ethical behaviorism requires that the robot’s actions are consistent with its utterances. Thus, if a robot declares that it “cares about you a great deal and wants to be of help” then its subsequent behavior should not be to avoid or ignore the user . Whilst it is easy to program a robot to make these kinds of supportive declarations it is much more difficult to make its behavior consistent with them. For instance, to be genuinely helpful, the robot must be able to recognize individuals consistently, perhaps remembering past encounters, and be able to monitor and anticipate the person’s needs, at least to some degree. Few, if any social robots, are capable of this level of helpful behavior at present [30]. On ethical behaviorism grounds, we might consider that the robot’s statement that it “cares” and “wants” to help as problematic to the extent that it raises expectations about its wider behavior that cannot be met, however, a future, more care-capable robot might more reasonably make such statements. As a further example of honest anthropomorphism we suggest that robots could have the ability of robots to track and recognize human emotions, and to modulate their own emotional expressions to be aligned with those of their human interlocutor [31]. People seek interactions in which their sense of self is respected and valued on an emotional level, alignment with artificial emotions could help to create this experience; moreover, AI technologies for emotion recognition are at the point of this being technologically feasible [31].
3. **Clearly signal the robot capacities**: The requirement to avoid hidden state deception suggests the importance of *clear signaling*. Here anthropomorphism can have some direct benefits, for example, if the robot’s only cameras are mounted forward-facing on its head, and can be covered by opaque eyelids, then closing the eyelids, or turning the head away, will be sufficient to communicate that the robot can no longer observe you. This is an intuitive and easy-to-read signal that matches our experience and expectations from interactions with people and pet animals. On the other hand, if the robot has other cameras, in anthropomorphically unexpected places (e.g. a rear-facing camera on a humanoid), then their presence/use should be very clearly signaled—for example, it has become conventional for cameras on computers to illuminate a small pilot light when they are operating. Dynamic feedback—emitting signals when the context changes—is likely to be important. For example, a home robot might usefully signal a switch from standby mode to awake/monitoring mode to alert users that its sensors have become operational.
4. Note that honest signaling is not the same as “transparency”, at least as that term has been used by Wortham [16] and others to imply transparency about the internal processes of the robot that underlie its decision-making etc. Signaling is here intended to avoid hidden-state deception and is not about revealing the robot’s machine nature. Of course, if the robot is asked about its internal processes it should answer

honestly (to the extent that it is capable), as to do otherwise would contravene broader principles around truth-telling and deception (See Danaher [17] for further discussion on this).

5. **Be especially careful when designing for vulnerable users and/or for “thick” relationships** (i.e. longer-term interactions with deeper psychological involvement). In assessing the potential benefits and risks, the relational approach emphasizes the need to consider the role of the robot within the wider network of the user’s interpersonal relationships. Social robots are currently developed and implemented in populations typically considered vulnerable, such as children with autism or with mental health conditions, and older adults living with dementia. These populations may be less able to make sophisticated judgments about meaning and intentions in social interactions. Ethical risks can be addressed through appropriate consent procedures involving family and carers, monitoring, and through careful co-creation of robot capabilities in order that these are aligned with the values of end-users. Where there is deeper psychological involvement there is also more risk of harm, but also the potential of greater benefit from providing robots with richer set of social capabilities.

4 Conclusion

In this paper we have sought to outline some considerations for the design of future social robots based on a relational ethics approach. We have sought to distinguish this from approaches predicated on a more essentialist (or substantialist) view that emphasizes ontological differences between human machines. Some of the latter approaches have argued that sociality in robots is wrong in principle, and that anthropomorphic features such as the ability to convey emotional signals are deceptive. Against this, we have argued that sociality can be a desirable and valued capability and that anthropomorphic features should be evaluated according to their risks and benefits. Benefits include ease-of-use and intelligibility for people. For instance, in persons living with Alzheimer’s disease, there is evidence that emotional processing is more resistant to decline than cognitive processing [32]. In seeking to eliminate aspects of interaction that carry emotional connotations, there is a risk that this could make otherwise beneficial technologies less engaging and therefore reduce adoption. More broadly, the relational approach emphasizes the need to consider the social setting and relationship context in which a robot is deployed, and the alignment of its behavior with prevailing norms. This argues for a pragmatic and inclusive approach to the design of assistive social robots, that involves potentials users and other stakeholders, in evaluating when and how social capabilities and anthropomorphic features can be safely and beneficially deployed.

Acknowledgement. This research was supported by the Wellcome Trust through “Imagining Technologies for Disability Futures” (ITDF) project.

References

1. Matarić MJ, Scassellati B. Socially assistive robotics. In: Siciliano B, Khatib O, editors. Springer Handbook of Robotics. Cham: Springer International Publishing; 2016. p. 1973-94.
2. Kabacińska K, Prescott TJ, Robillard JM. Socially assistive robots as mental health interventions for children: A scoping review. International Journal of Social Robotics. 2020. doi: 10.1007/s12369-020-00679-0.
3. Prescott TJ, Robillard JM. Are friends electric? The benefits and risks of human-robot relationships. iScience. 2021;24(1). doi: 10.1016/j.isci.2020.101993.
4. Coeckelbergh M. Robot rights? Towards a social-relational justification of moral consideration. Ethics and Information Technology. 2010;12(3):209-21. doi: 10.1007/s10676-010-9235-5.
5. Gunkel D. Robot Rights. Cambridge, MA: MIT Press; 2018.
6. Danaher J. Welcoming robots into the moral circle: A defence of ethical behaviourism. Science and Engineering Ethics. 2020;26(4):2023-49. doi: 10.1007/s11948-019-00119-x.
7. Prescott TJ. Robots are not just tools. Connection Science. 2017;29(2):142-9. doi: 10.1080/09540091.2017.1279125.
8. Emirbayer M. Manifesto for a Relational Sociology. American Journal of Sociology. 1997;103(2):281-317. doi: 10.1086/231209.
9. Szollosy M. EPSRC Principles of Robotics: defending an obsolete hu-man(ism)? Connection Science. 2017;29(2):150-9.
10. Haraway DJ. Simians, Cyborgs and Women: the Reinvention of Nature. New York: Routledge; 1991.
11. Hayles NK. How we became posthuman: Virtual bodies in cybernetifcs, literature, and informatics. Chicago: University of Chicago Press; 1999.
12. De Jaegher H, Di Paolo E, Gallagher S. Can social interaction constitute social cognition? Trends in Cognitive Sciences. 2010;14(10):441-7. doi: 10.1016/j.tics.2010.06.009.
13. Gunkel DJ. The Relational Turn: Third Wave HCI and Phenomenology. In: Filimowicz M, Tzankova V, editors. New Directions in Third Wave Human-Computer Interaction: Volume 1 - Technologies. Cham: Springer International Publishing; 2018. p. 11-24.
14. Sparrow R. The march of the robot dogs. Ethics and Information Technology. 2002;4(4):305-18. doi: 10.1023/A:1021386708994.
15. Bryson JJ. Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics and Information Technology. 2018;20(1):15-26. doi: 10.1007/s10676-018-9448-6.
16. Wortham RH, Theodorou A. Robot transparency, trust and utility. Connection Science. 2017;29(3):242-8. doi: 10.1080/09540091.2017.1313816.
17. Danaher J. Robot Betrayal: a guide to the ethics of robotic deception. Ethics and Information Technology. 2020;22(2):117-28. doi: 10.1007/s10676-019-09520-3.
18. Turkle S. Alone Together: Why We Expect More from Technology and Less from Each Other. 3rd ed. New York: Basic Books; 2017.
19. Boden M, Bryson J, Caldwell D, Dautenhahn K, Edwards L, Kember S, et al. Principles of robotics: regulating robots in the real world. Connection Science. 2017;29(2):124-9. doi: 10.1080/09540091.2016.1271400.
20. Shim J, Arkin RC. Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI? In: Agah A, Cabibihan J-J, Howard AM, Salichs MA, He H, editors. Social Robotics. Cham: Springer International Publishing; 2016. p. 222-32.
21. Sorell T, Draper H. Second thoughts about privacy, Safety and deception. Connection Science. 2017;29(3):217–22.

22. Danaher J. The philosophical case for robot friendship. Journal of Posthuman Studies. 2019;3(1):5-24. doi: 10.5325/jpoststud.3.1.0005.
23. Robillard JM, Goldman IP, Prescott TJ, Michaud F. Addressing the Ethics of Telepresence Applications Through End-User Engagement. Journal of Alzheimer's Disease. 2020;Preprint:1-4. doi: 10.3233/JAD-200154.
24. Damiano L, Dumouchel P. Anthropomorphism in Human–Robot Co-evolution. Frontiers in Psychology. 2018;9:468.
25. Heider F, Simmel M. An Experimental Study of Apparent Behavior. The American Journal of Psychology. 1944;57(2):243-59. doi: 10.2307/1416950.
26. Reeves B, Nass CI. The media equation: How people treat computers, television, and new media like real people and places. The Media Equation: How People Treat Computers, Television, and New Media Like Real pPeople and Places. New York, NY, US: CUP; 1996.
27. Kaminski ME, Rueben M, Smart WD, Grimm CM. Averting Robot Eyes Symposium Essays from the State of Cyberlaw: Security and Privacy in the Digital Age. Maryland Law Review. 2016;76(4):983-1024.
28. Leong B, Selinger E. Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism. Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA: Association for Computing Machinery; 2019. p. 299–308.
29. Nissenbaum H. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford, USA: Stanford Law Books; 2010.
30. Prescott TJ, Camilleri D, Martinez-Hernandez U, Damianou A, Lawrence ND. Memory and mental time travel in humans and social robots. Philosophical Transactions of the Royal Society B: Biological Sciences. 2019;374(1771):20180025. doi: doi:10.1098/rstb.2018.0025.
31. Robillard JM, Hoey J. Emotion and motivation in cognitive assistive technologies for dementia. Computer. 2018;51(3):24-34. doi: 10.1109/MC.2018.1731059.
32. König A, Francis LE, Joshi J, Robillard JM, Hoey J. Qualitative study of affective identities in dementia patients for the design of cognitive assistive technologies. Journal of Rehabilitation and Assistive Technologies Engineering. 2017;4:2055668316685038.