ICRES 2019: International Conference on Robot
Ethics and Standards, London, UK, 29-30 July 2019.
https://doi.org/10.13180/icres.2019.29-30.07.006

# THE NEED FOR ETHICAL PRINCIPLES AND GUIDELINES IN SOCIAL ROBOTS

A. VAN MARIS*, N. ZOOK, M. STUDLEY and S. DOGRAMADZI

*Bristol Robotics Laboratory,*
*University of the West of England,*
*BS16 1QY, Bristol, United Kingdom*
**E-mail: anouk.vanmaris@uwe.ac.uk*

This paper stresses the importance of establishing ethical principles regarding possible (psychological) effects of interactions with (social) robots. It highlights existing principles and standards regarding artificial intelligence systems and robots, and why these are not fully applicable to social robots yet. Lastly, it provides suggestions on how to establish ethical human-robot interactions.

## 1. Introduction

Research in social robots is ever increasing, and it is likely that these robots will become part of our everyday lives in the not so far future. However, with the increase in research in these topics, the awareness of ethical concerns has increased as well. Researchers have been developing standards for artificial intelligence systems, but can be applied to robots as well. One concern that is possibly less of a priority for artificial intelligence systems but extremely important for social robots is the (psychological) impact that (verbal) interactions with a social robot can have on its user. An essential difference between a robot and a social robot is that interaction is very important for the latter,[1] and therefore specific guidelines regarding this interaction are required. This paper discusses the existing artificial intelligence standards and identifies further requirements for social robot standards. It stresses the importance of standards for system-user dialogues and provides suggestions that may result in ethically acceptable human-robot interactions.

## 2. Ethical Concerns of Interactions Between A System and Its Users

People react to computers as social actors.[2] Therefore, they can become emotionally attached to social robots as well.[3] If emotional attachment to a robot is high, the usability of this robot is perceived more positively and the intention to use it in the future is higher.[4] As this is good for marketing, robots may be developed such that they elicit attachment. This can be done through showing emotions and/or deceptive behaviour, as people interact with social interfaces similar to how they interact with other humans.[2] Emotional attachment to a robot can be beneficial, as one can benefit from its assistive features more often. However, there are downsides to emotional attachment to a robot as well, as one can become (too) dependent on it, and its behaviour can have a psychological impact on the user. Providing behaviour that elicits attachment, like emotions or deception, can result in the user misunderstanding the abilities of the robot and therefore over-trusting it. This trust can be breached if at some point the robot does not meet the user's expectations. There may also be signs of withdrawal if the robot has a malfunction or is taken away. Lastly, users may inconvenience themselves by doing something they think robot requires due to its social behaviour.[5] As mentioned before, marketing may be one reason to implement unethical robot behaviour. It is possible that users are provided with sufficient information that their robot is deceptive or misrepresenting its internal state to increase usage (or other

reasons), but the result is still unethical. One might argue this was the case when social robot Jibo notified its users that the servers were going to be switched off[a]. Jibo stated that the time spent with its user(s) had been a pleasure. Users and people closely involved with the development of the robot knew that this was a goodbye message from the developers. However, this final notification of the robot was video-recorded and distributed on social media many times. People receiving these recordings do not have the information that users have. As a result, they may have an incorrect interpretation of the robot's internal state and adapt their expectations for future human-robot interactions based on this experience. Another cause of unethical robot behaviour can be from developers that do not realize the behaviour they are implementing is unethical. An example case scenario: a researcher is developing a learning algorithm that will improve interactions between the robot and its user, and they want to test this algorithm. The content of the interaction is not important for the test. It is possible that they implement an interaction that is similar to interactions between humans. However, this level of interaction requires an understanding of context that is often not feasible. Therefore, the participants interacting with the robots during the experiment experience an incorrect representation of the robot's abilities.

The cases described above emphasize the need for guidelines on the implementation of human-robot interactions. Seemingly innocent interactions between robots and their users can have consequences that developers did not foresee. Therefore, the remainder of this paper will discuss existing standards for artificial intelligence systems and where they potentially fall short regarding interactions with social robots, followed by suggestions that can initiate the development of guidelines and principles for ethical human-robot interaction.

## 3. Existing Standards and Principles

Several standards and principles have been established so far, both for artificial intelligence systems and robots.[6] However, the latter applies to robots in general and not specifically social robots, which means the vocal interaction has become less of a priority as many robots require no or minimal vocal interactions (e.g. war robots, drones). Existing principles and how they may not be fully applicable to social robots yet will be discussed below.

### 3.1. *Principles of Robotics: Regulating Robots in the Real World*

In 2010, a group of researchers developed five principles on how to regulate robots in the real world.[7] Most of these principles can be applied to social robots without the need for clarifications or additions. However, this does not hold for one of the principles, which entails emotional and deceptive social robot behaviour, as there are some ambiguities in the interpretation of this principle. It states illusion of intent and emotions (deceptive behaviour) are not to be used to exploit vulnerable users. However, this principle is difficult to interpret, as it does not specify what deceptive robot behaviour is, or how it will be determined whether a user is vulnerable (and by whom).[8] One may even take this further and claim that there should be no distinction in human-robot interaction for vulnerable people at all, as all interactions should be developed with extreme care. One cannot assume vulnerability, or vice versa, so all designs should minimize potential damage. Therefore, without going into detail on what can be defined as deceptive robot behaviour and whether this is acceptable or not, emotion and intent should not be used to exploit *any* user. Note that this argument tries to emphasize the need to be extremely careful with the development of social robot interactions, not demote awareness for extra needs and requirements that vulnerable users may have.

[a] https://www.bbc.co.uk/news/technology-47454599

### 3.2. *Robots and Robotics Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems*

The British Standards Institution published a first edition of a new standard that focuses on addressing issues that arise through new technologies, also known as BS 8611.[6] Ethical issues are divided into several groups, one of which defines societal issues. Ethical hazards of societal issues entail deception (either intentional or unintentional) and anthropomorphism. Unlike other ethical guides, BS 8611 notably acknowledges the concerns, but additionally provides possible approaches to soften the consequences of these concerns. They also provide societal ethical guidelines but, even though there is a guideline regarding deception, there does not seem to be one for anthropomorphism. Where BS 8611 shows reservations regarding anthropomorphism and anthropomorphic framing (personifying the robot by providing it with a name and background story etc.), others think that there is no reason for concern as long as the intended function of the robot is not affected.[9] However, emotional attachment to a robot results in people attributing names, personality and gender to the robot.[10] Therefore, it is questionable whether anthropomorphic framing elicits attachment, and whether this is ethically acceptable.

### 3.3. *Ethics Guidelines for Trustworthy AI*

In June 2019, the High-Level Expert Group on Artificial Intelligence published a draft containing ethics guidelines for trustworthy AI[b]. This document states that a human-centered approach to artificial intelligence is necessary to ensure the benefits of AI are maximized while the risks are minimized. Several principles are listed that will help developing AI in a human-centered way, one of them discussing the need for transparent processes. The document also provides an assessment list to assess trustworthy AI. In this list, the societal and environmental well-being are considered, which involves the interaction between the system and its user. There are assessments measuring whether it is clear to the user that any output from the system (a decision, advice etc.) is the result of an algorithmic decision, and that interactions are simulated, meaning the system has no understanding of feelings and emotions. However, no examples are provided regarding what (social) behaviour can for example elicit attachment. Unfortunately, the development of these behaviours that elicit attachment can be unintentional, and therefore go unnoticed. Examples or more specific guidelines would help to prevent this.

### 3.4. *Other Research Raising Awareness for Social Robot Ethics*

There are no generally accepted guidelines on what exactly ethical behaviour by a robot entails.[11] Although reaching consensus on this matter will be hard, it is necessary to have some guidance.[11] Some researchers raise awareness to the fact that users will become emotionally attached to the robot,[12] as also highlighted in this paper. However, they do not discuss this any further. As discussed before, emotional attachment to a robot has benefits, as the robot will be used more often and the user can benefit from its assistive features more. However, if users are emotionally attached to their robot, they can also become dependent on it, or be mentally stressed when the robot breaks down or is taken away.[12]

The Open RoboEthics Initiative focuses on the implementation of human ethics into social robot behaviours.[13] They demonstrate a use case for acceptable robot behaviour, in which the vocal interaction of the robot does not contain any of the concerns shown in the cases mentioned earlier in this paper, as it only states facts or decisions it has made without

[b] https://ec.europa.eu/futurium/en/ai-alliance-consultation

implying it would understand any underlying context of the situation. The study investigated (non-) yielding behaviour of a robot when riding an elevator. They varied in task urgency (urgent/not urgent) and an interacting person's location (riding/waiting for the elevator). Depending on the situation, the robot's (non-) yielding behaviour would differ. Besides verbal communication the robot also used body language to communicate its intent. The non-yielding behaviour that was used with the interacting person riding the elevator, entailed the robot exhibiting an arm gesture described as 'showing the person out of the elevator'. One might argue that this gesture is a form of body language, which is useful as it provides the robot with different means to communicate with its user. However, for laypeople with no experience with robots it may appear as if the robot has an understanding of politeness, giving them an incorrect representation of the robot's internal state and abilities. It may have been more transparent if the robot would have moved back to make room for the person to leave the elevator while stating *'I will wait for you to leave the elevator'*. It is essential to ensure that users fully understand the robot's abilities now that robots are entering our daily lives, as integration will be more successful and acceptance will be higher.

Other researchers discuss human dignity considerations (e.g. the emotional needs of humans should be respected) and social considerations (e.g. consideration of the fact that humans tend to form attachments to and anthropomorphize artificial systems).[14] However, these discussions do not provide suggestions on how these considerations should be addressed.

## 4. Suggestions and Guidelines on Social Interaction Standards

The standards, principles and concerns described above show that more detailed guidelines are required regarding the aspects of human-robot interaction that can have a psychological effect on the users (e.g. behaviour that elicits attachment, deception, or anthropomorphic framing). Building on the works described above to make them suitable for social robots can be a good way to start. It has also been established that transparency during human-robot interaction is essential. An initial principle with suggested guidelines that can hopefully be developed into standards and principles that can guide ethical human-robot interaction will be discussed next. The suggested principle is: **The psychological effect of an interaction between a social robot and its user(s) has to be ethically acceptable for all situations.** This applies to robots used both at home and in research. The latter is essential, as it cannot be expected that companies will follow ethical guidelines as no appropriate standard is provided by research. This indicates that researchers and companies need to carefully consider each interaction they develop, even if the intention of these interactions does not involve psychological influence on the users.

The final message of social robot Jibo was not transparent, as it gave the impression that Jibo had an understanding and feelings about the servers being shut down. Therefore, the first suggested guideline for ethical human-robot interaction is the following: ***Social robots should only provide useful information and not give the impression they have personal opinions.*** Providing personal opinions gives the user the impression that the robot has a full understanding of context, which it does not (yet) have. The best way to apply this suggestion is by ensuring the robot does not phrase sentences with *'I'*, unless it is in an informative setting (e.g. *'I will bring you a glass of water'*, but not *'I think you are right'*). One might argue that interactions will become too static, which can impact the use of the robot, but this does not have to be true. If a user asks a question to which the robot cannot give an ethically acceptable answer, it can reply that it cannot answer the question but provide other options that might be useful for the user.

The second suggestion builds on principle four from 'Principles of Robotics: Regulating Robots in the Real World', that states that the illusion of emotion and intent should not be used to exploit (vulnerable) users. This suggestion is the following: ***The use of anthropomorphic and/or deceptive abilities has to be ethically validated before it is implemented in the social robot.*** This does not imply that they should not be used, as research has shown that they increase successful human-robot interaction, which is important for trust, acceptance and use of the robot.[15] However, the use of a deceptive or anthropomorphic feature should always be validated to be ethically acceptable. One might suggest that user studies are required to validate this, which would mean the user studies are possibly unethical. However, risk assessment regarding the psychological effect can already be a good start. Founding groups that specifically investigate whether interactions are ethically acceptable may be useful, similar to the ethics committees that currently have to give approval for user studies (but often do not focus on the interaction itself). Also, the interaction should be developed with the mindset that all users are vulnerable. Additional requirements can be added for people from vulnerable user groups.

The next suggestion is the following: ***Social robot behaviour should not mimic human behaviour.*** Social robots are beneficial and can be useful additions help people with their tasks, but this does not mean they should replace people and behave like them. For every interaction of the robot, it has to be considered whether it is true to the robot's abilities or whether it is something a human would say during human-human interaction. If the latter is the case, the interaction should possibly be rephrased.

The final suggestion is: ***In interactions where the internal state and/or intention of the social robot are ambiguous, it should announce that it is a machine.*** This is adapted from Turings Red Flag law[16] such that it applies to human-robot interactions. Ambiguous interactions can occur when the robot has to provide information about its internal state, or when a user would ask a question that the robot cannot or should not answer.

## 5. Conclusion

This paper aimed to stress the necessity for standards and principles regarding the social interaction between a social robot and its user. Existing standards and principles for robots and artificial intelligence systems have been discussed, and suggestions have been provided on how they can be extended to also apply to social robots. The aim for the future is to develop these with the help of researchers from a variety of research areas, and publish them as official standards and principles. Hopefully, this will help to establish a future where robots provide only benefits, as other technologies already make everyday life and decision making difficult enough.

## Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

## References

1. T. Fong, I. Nourbakhsh and K. Dautenhahn, *Robotics and autonomous systems* **42**, 143 (2003).
2. B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places.* (Cambridge university press, 1996).
3. J. P. Sullins, *IEEE transactions on affective computing* , p. 1 (2012).
4. C. Wilson, *Ageing & Society* , 1 (2017).

5. T. A. Garner, W. A. Powell and V. Carr, *Digital health* **2**, p. 2055207616681173 (2016).
6. BS. 8611:2016, *London:British Standards Institution* (2016).
7. M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden *et al.*, *Connection Science* **29**, 124 (2017).
8. E. C. Collins, *Connection Science* **29**, 223 (2017).
9. K. Darling, *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015). ROBOT ETHICS* **2** (2015).
10. J.-Y. Sung, L. Guo, R. E. Grinter and H. I. Christensen, my roomba is rambo: intimate home appliances, in *International Conference on Ubiquitous Computing*, 2007.
11. J. Borenstein and R. Arkin, *Science and engineering ethics* **22**, 31 (2016).
12. G. Veruggio, F. Operto and G. Bekey, Roboethics: Social and ethical implications, in *Springer handbook of robotics*, (Springer, 2016) pp. 2135–2160.
13. E. Calisgan, A. Moon, C. Bassani, F. Ferreira, F. Operto, G. Veruggio, E. Croft and H. M. Van der Loos, Open roboethics pilot: accelerating policy design implementation and demonstration of socially acceptable robot behaviours.
14. L. Riek and D. Howard, *Proceedings of We Robot* (2014).
15. A. Waytz, J. Heafner and N. Epley, *Journal of Experimental Social Psychology* **52**, 113 (2014).
16. T. Walsh, *arXiv preprint arXiv:1510.09033* (2015).