
Article:
Sharkey, A. and Sharkey, N. (2021) We need to talk about deception in social robotics! Ethics and Information Technology, 23 (3). pp. 309-316. ISSN 1388-1957

https://doi.org/10.1007/s10676-020-09573-9

Ethics and Information Technology (2021) 23:309–316
https://doi.org/10.1007/s10676-020-09573-9

ORIGINAL PAPER

We need to talk about deception in social robotics!

Amanda Sharkey[1] · **Noel Sharkey**[1]

Accepted: 29 October 2020 / Published online: 11 November 2020

Abstract
Although some authors claim that deception requires intention, we argue that there can be deception in social robotics, whether or not it is intended. By focusing on the deceived rather than the deceiver, we propose that false beliefs can be created in the absence of intention. Supporting evidence is found in both human and animal examples. Instead of assuming that deception is wrong only when carried out to benefit the deceiver, we propose that deception in social robotics is wrong when it leads to harmful impacts on individuals and society. The appearance and behaviour of a robot can lead to an overestimation of its functionality or to an illusion of sentience or cognition that can promote misplaced trust and inappropriate uses such as care and companionship of the vulnerable. We consider the allocation of responsibility for harmful deception. Finally, we make the suggestion that harmful impacts could be prevented by legislation, and by the development of an assessment framework for sensitive robot applications.

Keywords Robot · Deception · Intentional deception · Harm · Robotics · False belief · Prevention · Social robotics · Illusion

"Most of the evil in this world is done by people with good intentions."— T.S. Eliot.

Introduction

According to a number of authors (e.g. Matthias 2015; Sparrow and Sparrow 2006; Sparrow 2002; Wallach and Allen 2009; Sharkey and Sharkey 2011), the development and creation of social robots often involves deception. By contrast, some have expressed doubts about the prevalence of deception in robotics (e.g. Collins 2017; Sorell and Draper 2017). It seems that there is disagreement in the field about what counts as deception, and whether or when it should be avoided.

The 4th principle of the U.K. Engineering and Physical Sciences Research Council's (EPSRC) (Boden et al. 2017) 'principles of robotics' states that 'Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be made transparent'. Although this principle is a step in the right direction, there is a need for a more detailed consideration of what constitutes deception in social robotics, when it is wrong, who should be held responsible, and whether it can be prevented or avoided.

A social robot is a physically embodied robot that is able to socially interact with people. Wallach and Allen (2009) hold that any techniques enabling robots to detect basic human social gestures and to respond with human-like social cues, "are arguably forms of deception" (pp 44). Matthias (2015) suggests that a robot that appears to have mental or emotional capabilities that it does not really have, is implicated "in a kind of deception" (pp 17). Grodzinsky et al. (2015) declare that deception is involved if the behaviour of a machine "leads another agent to believe or behave as if the machine is human or some other carbon-based life form". Sharkey and Sharkey (2011) see "efforts to develop features that promote the illusion of mental life in robots as forms of deception", since current robots have neither minds nor experiences (pp 34). Johnson and Verdicchio (2018) consider the possible appearance of suffering in either animals or robots and conclude that "both [can] involve the appearance of suffering but when it comes to robots, the appearance is a deception" (pp 299).

Sorell and Draper (2017) are more sceptical about deception in robotics. They set a high bar and assume deception is involved only if the design of a robot misleads people into

✉ Amanda Sharkey
a.sharkey@shef.ac.uk

[1] Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

believing that it is a real human or animal. This bar would not often be surmounted, since it is rarely the case that people are fooled into believing that a robot is actually a human, or an animal (unless it is viewed under quite limited conditions, such as for a very short time period—see Ishiguro 2006). Collins (2017) also argues that animal-like robots are not developed with the intention of convincing the user that the robot is alive.

A primary aim of this paper is to make the case that, despite some claims to the contrary, there is deception in social robotics. The idea that there is deception is sometimes rejected on the basis that the developers and programmers of social robots may not have intended to deceive. In an effort to counter this resistance, we examine definitions of deception, and argue that deception can still occur even in the absence of conscious intention. If a person believes that a social robot has emotions and cares about them, they are being deceived: even if no-one explicitly intended that belief. An important first step to preventing the harmful consequences of deception in social robotics is to recognise that it can and does occur. Once this is recognised, an important next step is to consider what harmful effects it could have. We identify potential examples of such effects and follow this with a discussion of who could be held responsible, and whether such consequences could be prevented.

Deception without intention

Underlying some of the disagreement about whether or not deception is involved in social robotics are differences of opinion about whether it necessarily involves intention on the part of the deceiver. Sorell and Draper (2017) propose that it requires "the intentional creation of false beliefs". They argue that other authors (Sharkey and Sharkey 2011) set too low a threshold for deception by not requiring the creation of an intentionally produced false belief that a robot is alive. As an example, they consider the Paro, a fur covered seal-like robot used for therapy for older people. They argue that no deception is involved in its use, since it can provide comfort without the need for an intentionally produced false belief that the robot is a real seal.

We disagree and argue that there can be deception without intention. In the case of Paro, the manufacturers did not intend to create the false belief that the robot is an actual seal. Rather they wanted to create a therapeutic object that reminded people of a pet and gave them some relief from stress. Nonetheless, the illusion of sentience or cognition created for some people by its appearance and behaviour can be said to be a deception.

Several other authors define deception as involving intention. For instance, Carson (2010) 'roughly' defines deception as 'intentionally causing someone to have false beliefs' ('roughly' because he has more detailed definitions associated with particular cases). Carson points out a difference between lying and deception in that unlike lying, 'deception' implies success: in order to count as deception, an act must actually cause someone to have false beliefs. Another definition of deception is provided by Zuckerman et al. (1981) and requires that a human be deceived: 'an act that is intended to foster in *another* person a belief or understanding which the believer considers to be false'. Grodzinsky et al. (2015) say that their main interest is represented by examples of 'an intentional, successful attempt by developers to deceive users'.

Bok (1999) makes a distinction between deception and lying. For her, lying does involve intention, but the same is not necessarily the case for deception. In her book on 'Lying', Bok defines lying as 'any intentionally deceptive message which is stated' (Bok 1999). However, she claims that when a person states false information that they believe at the time to be true, it should not be described as a lie even though it engenders a deception, for any deception that occurs as a result was not intended by them. Bok identifies various situations in which people might convey false information in the belief that it is true. These include cases where they have been misinformed, where they are mistaken, or tired, or inarticulate or even delusional. In situations like these, deception may have occurred even though there was no intention to deceive.

We agree with Bok that intention is not a necessary condition for a deception to have occurred. This can be seen more clearly when perspective is shifted from the deceiver to the deceived. For example, Fallis and Lewis (2019) refer to deception in the animal kingdom as 'functional deception', recognising that it involves evolution rather than intention.

Discussing the evolution of deception, Bond and Robinson (1988) define it as 'a false communication that tends to benefit the communicator'. Examples include camouflage, mimicry, death feigning, and distraction displays:

- *Camouflage* confers an evolutionary advantage to the peppered moth, *Biston betularia*. It comes in two colours, speckled and black, and in polluted areas the black form has come to dominate making the moth less visible to predators.
- *Mimicry* can be found in the viceroy butterfly, which is palatable, but which has the markings of a monarch butterfly that birds recognise as inedible.
- *Death feigning*, a deception and anti-predator adaptation, occurs in a range of animals: lizards, birds, rodents and sharks (Pasteur 1982).
- *Distraction displays*, that draw attention away from nests and young, are found in both birds and fish (Armstrong 1954).

We also don't have to look far to find examples of deception without intention in the human world: for instance, the doll therapy used for people with dementia (Sharkey and Sharkey 2010b). Off-the-shelf baby dolls are used to therapeutically stimulate memories of a rewarding life role, and to act as a focus for reminiscence and conversation (Cayton 2006). James et al. (2006) point out that even though these dolls were designed as children's toys, many clients come to believe that their doll is a real baby. They try to feed it and are even prepared to give up their own bed for it to have a good night's sleep (ibid). Likewise, there are men who believe that their static sex doll is in love with them and that they are in a loving relationship with it (Sharkey et al. 2017). This was probably not intended by the developers of sex dolls.

These examples show that both non-human animals and humans can be subject to deception without any intentional attempt to deceive. They support our claim that a deception can be said to have occurred in robotics if the appearance and the way that a robot is programmed to behave, creates, for example, the illusion that a robot is sentient, emotional, and caring or that it understands you or loves you. These are all deceptions because, at present and for the foreseeable future, robots are machines that are not alive, sentient or capable of feelings. Yet none of these deceptions necessarily require a conscious intention on behalf of robot manufacturers or designers who may simply have meant to entertain.

When is deception wrong?

Some deceptions can be harmless fun. For example, show robots used in the leisure industry can entertain and motivate the next generation of engineers[1]. There can also be situations in which people are both entertained and aware that the illusion of sentience created by a social robot is not real. Coeckelberg (2018) argues that any deception or illusion created by information technology is the result of a performance 'co-created and co-performed by humans (magician/designer and spectator/user) and non-humans (robots and other machines, artefacts and devices)' (Coeckelberg 2018, pp 78). As in the case of a magic show, the users/audience may not have been fooled into thinking that a robot is sentient, or has emotions, and may know that it is a trick. This is likely to be the case when a social robot is displayed to an audience with a reasonable knowledge of robotics. The audience members could enjoy the performance, at the same time as looking for clues or asking questions about how the performance was accomplished. This is less likely to be the case with naïve audiences, or vulnerable groups of people such as the very young, or older people with cognitive limitations.

As well as entertainment, some deceptions can be created with good intentions that lead to better outcomes for the deceived. Bok (1999) gives several examples of deceptions created with the good intention of helping or protecting the deceived, including placebos, white lies, or the lies that used to be told to the sick and dying (a practice now uncommon). And in robotics, there is evidence that animated robot pets, such as the Paro, can improve the health and well-being of vulnerable users (Robinson et al. 2013).

At the other extreme, Sparrow (2002) argues that the self-deception involved in an imaginary relationship with a robot is inherently wrong and violates a duty to see the world as it is. Presumably his position would also apply to the treatment of people with dementia, even though some deceptions (such as offering them a robot seal pet to care for) might alleviate their distress and anxiety. He also argues that those who design and manufacture robots that encourage such beliefs are unethical.

In contrast, as mentioned previously, Sorell and Draper (2017) focus on the intentions of the deceiver. So, for them, deception in robotics is wrong only when the deceiver wants to manipulate the deceived person to do something that serves the interests of the deceiver. In other words, when the deceiver has malign, or at least self-serving, intentions.

Our argument is that determining whether or not deception in robotics is wrong should be based on assessments of the likely impact on individuals and society regardless of the beneficent, or malignant, intentions of a deceiver. Efforts to anticipate the possible negative effects and risks of robotics applications underlies much of the emerging field of robot ethics. It is important to make the attempt to foresee risks and concerns before the applications become so deeply entrenched in society that they can no longer be prevented or even curtailed.

It is possible to identify two kinds of risk that could result from the development and presentation of social robots that appear to have emotions and to be able to understand and care about humans. We consider in turn (i) those that stem from the deception involved in robots that appear to have emotions and to care for us; and (ii) those that originate in over-estimations of the ability of robots to understand human behaviour and social situations.

First, we consider the possible negative consequences of any emotional deception that encourages us to believe that a robot cares for us, has emotions, and is something with which we could form a relationship.

For young children and babies, there are clear risks to emotional deception. At the extreme, leaving babies in the 'care' of robots for prolonged periods could interfere with

[1] Although this can be problematic under some circumstances when it leads to general and harmful overestimation of the technology (see Sharkey, 2018).

the formation of secure attachments to their primary human carers (Sharkey and Sharkey 2010a). We can only speculate about the risks since it would clearly be unethical to conduct experiments on babies with such a risk of harm. We turn instead to indirect evidence that demonstrates the creation of attachment disorders such as the disturbing reports of children in Romanian orphanages deprived of human interaction (Nelson 2007; Chugani et al. 2001). And Harlow's classic experiments (Harlow and Zimmerman 1958), in which young rhesus monkeys were left in the exclusive company of artificial wire or cloth covered 'mothers'. These provide indicative evidence of the devastating effects of the loss of attachment and bonding opportunities with a significant living being.

Even without the extremes of full-time child care, there is still a considerable interest in developing robot companions and internet-connected toys for children. One risk of having such companions is that children who spend too much time interacting with them will miss out on opportunities to learn about the natural give and take of human relationships with their peers and might even come to prefer the predictability of a robot companion that always agrees with them, always listens to them, and always puts up with their selfish behaviour.

Turkle (2017) is also concerned about the idea of such companions and writes, 'These machines are seductive and offer the wrong payoff: the illusion of companionship without the demands of friendship, the illusion of connection without the reciprocity of a mutual relationship. And interacting with these empathy machines may get in the way of children's ability to develop a capacity for empathy themselves.' There is already a growing awareness of the addictive effect of technologies such as social networks, or mobile phones (Kuss 2017) and it seems that a physically embodied and rewarding robot companion would be similarly difficult to resist. As in the case of babies and robot carers, the potential risks raised here are inevitably speculative. Because research based on limiting children's companions to robots to the exclusion of human friends is unlikely to receive ethical approval, there is little clear evidence of the detrimental effect on children as a consequence of long-term interactions with robot companions.

Scheutz (2011) suggests another risk to people forming emotional uni-directional bonds with trusted social robots. A company could exploit such relationships to encourage the purchase of their products. And there is indeed evidence from Tanaka et al. (2007) that 'long-term bonding' occurred between toddlers and a social robot operated over 5 months in a day-care centre. Children who believe that a robot (or internet-connected toy such as 'Hello Barbie') is their friend might confide in it without any awareness of the ways in which their confidences might be shared or used to manipulate their purchasing behaviour. This is quite different to the bonds that children form with inanimate, unconnected, teddy bears and dolls.

There is evidence that adults form attachments to robots even when they are remote controlled bomb disposal robots (Carpenter 2016) or robot vacuum cleaners (Sung et al. 2007). When the robots are furry robot pets, or cute looking humanoids, attachment formation is even more likely. Such emotional attachments could have negative consequences for vulnerable adults such as those with dementia or other cognitive limitations (Sharkey and Sharkey 2012; Sharkey 2014). They might choose to neglect their relationships with fellow humans to focus their emotions and attentions on the robot instead. They could become anxious and concerned about their robot companions**.** In addition, robot companions which give rise to the deceptive illusion that they care and understand, could result in a reduction of contact with other human beings for vulnerable individuals. Friends, family, and care providers in general, might come to believe that the social and attachment needs of an older person were being met by a robot companion or pet, and as a consequence might reduce the time they spent with them.

As in the case of children, an older person who believed a robot to be their friend might share information with it that they would not like to be shared more widely. They might also follow its advice, which cannot be guaranteed to always be appropriate for their situation.

The second set of risks considered here stem from an overestimation of the ability of robots to understand the world that results in delegating decisions to them that impact on the quality of a human life. If the illusion is created by means of their appearance and behaviour that robots are able to understand our world, and to make justifiable decisions, a risk of such deception is that robots could be inappropriately deployed in social roles for which they are unsuited. Sharkey (2016) considers the use of robot teachers in classrooms, where they could be required to make decisions for which they are not equipped. These could be decisions about what counts as good or bad behaviour in children. Or decisions about a child's readiness to learn something. Although many teachers are clear that robots would not have the ability to replace humans in the classroom (Serholt et al. 2017), this might not be so clear to educational authorities, particularly when there are staff shortages or when budget cuts have to be made.

Robots used for the care of children, or fragile old people, might also be delegated with decision making that goes beyond their functionality. For instance, a robot looking after children might be expected to prevent them from injuring themselves—but there is a myriad of circumstances in child care that it would be impossible for programmers to anticipate. For example, would a robot be able to recognise the difference between a child picking up scissors for a craft project, and a child using scissors

to dangerously poke into a toaster? Of course, some dangerous situations could be anticipated by the robot's developer, but real life in the real world is crowded with unanticipated and unpredictable events that could lead to negative consequences.

The creation of a robot, or computational artefact, that encourages the belief that it is able to make moral decisions is particularly concerning. There is a growing awareness of the risks associated with algorithmic decision making by machines that are trained on big data, but which have no understanding of the meaning or effects of their decisions (Sharkey 2018). An extreme example of algorithmic decision making can be found in the example of lethal autonomous weapons, where concerns are being raised about giving robots the power to make life or death decisions about who to kill on the battlefield (see the Campaign to stop Killer Robots website for arguments for a ban of such weapons). An example that is closer to home is that of the autonomous car, and discussions of its decisions about where to turn and who to hit in the event of an accident (Lin 2016). The appearance and behaviour of a social robot that creates an illusion of understanding could foster belief in its moral competence.

In this section, we have considered two kinds of risks stemming from deception in social robots. There is a set that stem from imagining that a robot is emotional and able to care for humans. If babies and children were to be left in the 'care' of robots for long periods, there is a danger that this could have a serious effect on their social and emotional development. Children and vulnerable older people could turn away from human companions in the mistaken illusion that a robot was something they could have a relationship with. There is evidence (Epley 2007) that those who are lonely and in need of social contact are more likely to engage in anthropomorphism. Both young and old might also confide in the robot they think is their friend, without understanding that their confidences might be shared more widely. Erstwhile robot 'companions' might also be used to encourage or pressurise them to make unnecessary purchases.

There are also risks that arise from being deceived into overestimating the abilities of robots, and inappropriately placing them in positions of responsibility, e.g. in charge of classrooms, or of the vulnerable in care homes. People can overestimate the strengths and competences of a computational algorithm, but the development of robots with humanoid appearances that appear to understand and respond to human emotions and needs, has the potential to be that much more compelling.

There is a particular need to recognise the risks for the more vulnerable members of society, who are often dependent on others to make decisions for them. The youngest and oldest members have a greater need for protections to be in place to limit their exposure to robots, to ensure their access to human companionship and care and to prevent their exploitation by social robots.

Who is responsible, and can we prevent deception in social robotics?

It may be impossible, even undesirable, to prevent all deception in robotics. As we have seen, some deceptions can lead to health benefits. There are also humanoid, and animal-like, robots that are entertaining: anthropomorphic entertainment devices date back to antiquity.

However, where there is a risk that deception in social robotics could lead to harmful consequences, it would be wise to attempt to find ways of preventing or limiting it. It would be easier to prevent harmful deceptions if we could identify who should be held responsible. The problem is that establishing responsibility is difficult because, as argued earlier, it can occur in the absence of an 'intent to deceive'. The challenge is made greater by the number of people usually involved in the development of a robot application. But before discussing this further let us first lay to rest the idea that a robot itself could be held responsible for deception.

Even an autonomous robot acting without concurrent human control can do only what it has been programmed or trained to do. Robots are dependent on human intervention for their programming, or for setting up and designing the conditions for machine learning. Occasionally roboticists inappropriately describe their robots as being deceptive. For example, Arkin et al. (2012) reported a study in which he said that robots had deceived an observer about which route they had taken. But the deceptive behaviour of the robots was explicitly hand coded: the programmer was responsible for the deception and not the robots. Even when deceptive behaviour seems to emerge spontaneously, as in a case reported by Floreano et al. (2007), human intervention is still involved. In that study, robots were shown to evolve a capacity to send deceptive signals to other robots. However, this 'evolution' was dependent on the situation in which humans placed them, and the human programming of the evolutionary design and objective function.

If robots cannot be held responsible for deception, could the users of robots be implicated? It is well known that humans are anthropomorphic about robots and other machines, or any inanimate objects with certain features. It is less clear that humans have much choice in this matter. An animated and apparently needy robot can be hard to resist. As pointed out by Turkle (2011), a robot that seems to require care and nurturing is particularly compelling. Vulnerable old people and young children may be especially susceptible to such anthropomorphic cues (Epley 2007).

Even if humans may sometimes be willing collaborators in the illusions created by robots, responsibility can be still

be attributed, at least in part, to manufacturers and marketers. Those marketing or advertising a robot may exaggerate its abilities, as in descriptions of the Cozmo robot as having 'real emotions'. The designers and programmers may also exploit features that encourage the illusion of sentience and understanding. These include the robot's appearance, the facility to detect human emotions and to indicate an emotional response, and the facility for speech recognition and speech generation. Excuses and justifications for such features and any exaggerated descriptions may be based on claims about intended beneficial effects, and a denial of any intention to deceive.

Although it can be argued in some cases that there was no clear intent to deceive, robot developers, programmers and sellers should bear some of the responsibility if they should have foreseen a harmful deception. Matthias (2015) argues that if a machine says, 'I love you', deception is necessarily involved because the machine lacks the corresponding mental state. He argues that such deception can be 'foreseen but not intended'. When a robot pet is created to entertain a child, the humans involved in its development may have intended to amuse and educate rather than to deceive. But they should have been able to foresee children's misinterpretation of the robot's actions as indicating 'a genuine, alive pet.' This also applies to humanoid robot companions for either children, or older people. Although the developers might not intend any deception, it could be argued that they should have foreseen that such users would perceive the robot as having a meaningful relationship with them.

Unfortunately, even if it is accepted that the programmers, developers and marketers of social robots bear some responsibility for the deceptions they can engender, it is still difficult to see how to prevent harmful effects. Prevention is made harder by the human tendency to be fascinated by technology, and by people's willing and enthusiastic tendency to be anthropomorphic. Nonetheless, we can consider some suggestions.

Scheutz (2011) suggested two ways to minimise the negative effects of deception, but we do not think that either are likely to be effective. One suggestion is to legally require a robot to continuously remind the user that it is only a machine and has no emotions. But it is not clear that this would prevent people from forming an attachment to it, and it might interfere with the comfort and relief from anxiety that interactions with a robot pet may create for some people requiring it for therapeutic reasons. The second suggestion is to equip the robot with an emotional system or to develop robots with a sense of morality. This seems unlikely to happen in the near future. There have been attempts to program robots with ethical rules (Anderson et al. 2006; Winfield et al. 2014), but these are rules that lack universality. They apply only in very limited and highly specific contexts. For example, Anderson et al. 2006) programmed a robot to use a set of ethical rules to decide whether or not to remind a patient to take their medicine, and whether or not to report the patient for not taking their medicine. As argued by Sharkey (2017), there is as yet little indication that it will be possible to code a set of ethical rules that will work in all circumstances and in many contexts.

We suggest that a better approach would be to put the onus of proof on robot manufacturers and sellers. Just as in the pharmaceutical industry, they could be required to provide evidence that a given robot application would not cause psychological harm or derogate any human rights. Thus, certain sensitive applications such as the use of robots in care situations could have a default of prohibition unless convincing evidence was provided that demonstrated benefits to wellbeing. Sensitive applications would include those involving babies, young children, and vulnerable older people. Some form of quality mark could be established to indicate that the robot had passed a set of ethical checks. The Foundation for Responsible Robotics has recently carried out a pilot project on the development of an assessment framework for a quality mark for AI based robotics products (FRR pilot report, under embargo) The planned quality mark would assess robot products for the extent to which they adhered to a set of 8 principles for robots; namely (1) security, (2) safety, (3) privacy, (4) fairness, (5) sustainability, (6) accountability and (7) transparency and (8) well-being. The approach is promising and could be adapted to include consideration of the risks of deception in social robots as explored here.

There are other measures that could be put in place to limit the exploitation of users. One example is that legislation could be passed to prevent robots that masquerade as friends or companions from using users' data to manipulate them, for example, in purchasing advertised items. Another possible measure would be to require that any sharing of information obtained by the robot is made transparent and explicit. As elucidated by Zuboff (2019), there are powerful forces behind 'surveillance capitalism' that are likely to mount a strong resistance to such attempts, but that does not mean it cannot be done. Attention should also be paid to assessing and limiting promotional descriptions of robots, that exaggerate their functionality and their benefits.

Clearly these suggestions are merely a first step towards developing a comprehensive framework for the avoidance of harmful deceptions. Developing such a framework will be challenging because people can be entertained by and enthusiastic about interactive robots, but it is crucial to provide some means of protection for the more vulnerable members of society who may not have a good understanding of the robots' limitations.

Summary and Conclusions

In this paper, we argue for the need to recognise that there is deception in social robotics. Reflecting on its prevalence, we argued that intention is not necessary for deceit to have occurred. We garnered support from studies of humans and other animals. For example, vulnerable humans have been shown to be deceived by objects such as dolls that were intended for children's play. In the animal kingdom, predators can be deceived by camouflage, mimicry, death feigning, or distraction displays that are the result of evolution rather than deception. If the behaviour and appearance of a robot leads to people believing that a robot has cognitive abilities or that it cares for and loves them, then, we argue, they are being deceived whether or not anyone intended to deceive them.

We do not suggest that all deceptive illusions in robotics are wrong. Social robots can be entertaining and fun to interact with. Instead we argue that deception is wrong when it creates negative impacts on individuals and society. The potential harmful impact of deceptions that result in an overestimation of a robot's functionality include their inappropriate use to replace human care, and a misplaced trust in their ability to make decisions for which they are not qualified.

Responsibility for wrongful deception can be attributable to the combined contributions of users, developers and marketers of robot applications. Preventing harmful deception is difficult, but our suggestion is for an evidenced quality or kite mark indicating that tangible efforts have been made to foresee, recognise and avoid likely negative consequences and demonstrating any claimed benefits. It is important to find ways to ensure that deception in social robotics does not lead to robots replacing meaningful human care, or to misplaced trust in decisions made by machines.

References

Anderson, M., Anderson, S., & Armen. (2006). MedEthEx: A prototype medical ethics advisor. In *Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE, 100,* 571–589.

Armstrong, E. A. (1954). The ecology of distraction display. *The British Journal of Animal Behaviour*, *2*(4), 121–135.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science, 29*(2), 124–129.

Bok, S. (1999). *Lying: Moral choice in public and private life*. New York: Second Vintage Books Edition.

Bond, C. F., & Robinson, M. (1988). The evolution of deception. *Journal of Nonverbal Behavior, 12*(4), 295–307.

Carpenter, J. (2016). *Culture and human–robot interaction in militarized spaces: A war story*. New York: Ashgate.

Carson, T. L. (2010). *Lying and deception: Theory and practice*. New York: Oxford University Press Inc.

Cayton, H. (2006). From childhood to childhood? Autonomy and dependence through the ages of life. In J. C. Hughes, S. J. Louw, & S. R. Sabat (Eds.), *Dementia: Mind, meaning, and the person* (pp. 277–286). Oxford, UK: Oxford University Press.

Chugani, H., Behen, M., Muzik, O., Juhasz, C., Nagy, F., & Chugani, D. (2001). Local brain functional activity following early deprivation: A study of post-institutionalised Romanian orphans. *Neuroimage, 14*(6), 1290–1301.

Coeckelbergh, M. (2018). How to describe and evaluate "deception" phenomena: Recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology, 20,* 71–85.

Collins, E. C. (2017). Vulnerable users: Deceptive robotics. *Connection Science, 29*(3), 223–229. https://doi.org/10.1080/09540091.2016.1274959

Epley, N., Waytz, A., & Caciopo, T. (2007). On seeing human: A three factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

Fallis, D., & Lewis, P. J. (2019). Towards a formal analysis of deceptive signalling. *Synthese, 196,* 2279–2303.

Floreano, D., Mitri, S., Magnenat, S., & Keller, L. (2007). Evolutionary conditions for the emergence of communication in robots. *Current Biology, 17*(6), 514–519.

FRR pilot report (under embargo) Report on the Pilot Phase of the Foundation for Responsible Robotics Quality Mark Project.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2015). Developing automated deceptions and the impact on trust. *Philosophy & Technology, 28*(1), 91–105.

Harlow, H. F., & Zimmermann, R. R. (1958). The development of affective responsiveness in infant monkeys. *Proceedings of the American Philosophical Society, 102,* 501–509.

Ishiguro, H. (2006). Android science: Conscious and subconscious recognition. *Connection Science*, *18*(4), 319–332.

James, I. A., Mackenzie, L., & Mukaetova-Ladinska, E. (2006). Doll use in care homes for people with dementia. *International Journal of Geriatric Psychiatry, 21*(11), 1093–1098.

Johnson, D., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology, 20,* 291–302.

Kuss, D. (2017). Mobile phone addiction: Evidence from empirical research. *European Psychiatry, 41*(Supplement), S26-27.

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. Chrisian Gerdes, B. Lens, & H. Winner (Eds.), *Autonomous driving* (pp. 69–85). Berlin: Springer Open.

Matthias, A. (2015). Robot lies in health care: When is deception morally permissible? *Kennedy Institute of Ethics Journal, 25*(2), 169–162.

Nelson, C. A., Zeanah, C. H., Fox, N. A., Marshall, P. J., Smyke, A. T., & Guthrie, D. (2007). Cognitive recovery in socially deprived young children: The Bucharest early intervention project. *Science, 318*(5858), 1937–1940.

Pasteur, G. (1982). A classificatory review of mimicry systems. *Annual Review of Ecology and Systematics*, *13*(1), 169–199.

Robinson, H., Macdonald, B., Kerse, N., & Broadbent, E. (2013). The psychosocial effects of a companion robot: A randomized controlled trial. *Journal of the American Medical Directors Association, 14*(9), 661–667.

Scheutz, M. (2011). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, G. Bekey, & K. Abney (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 205–222). Cambridge, MA; London: MIT Press.

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., & Paiva, A. (2017). The case of classroom robots: teachers' deliberations on the ethical tensions. *AI and Society, 32*(4), 613–631.

Sharkey, A. (2014). Robots and human dignity: The effects of robot care on the dignity of older people. *Ethics and Information Technology, 16*(1), 53–75.

Sharkey, A. (2016). Should we welcome robot teachers? *Ethics and Information Technology, 18*(4), 283–297. https://doi.org/10.1007/s10676-016-9387-z

Sharkey, A. (2017). Can we program or train robots to be good? *Ethics and Information Technology, Online First*. https://doi.org/10.1007/s10676-017-9425-5

Sharkey, A., & Sharkey, N. (2011). Children, the elderly, and interactive robots. *IEEE Robotics and Automation Magazine, 18*(1), 32–38.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*(1), 27–40.

Sharkey, N. (2018) The impact of gender and race bias in *AI.* August 28,2018 https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/ Accessed December 6th 2018

Sharkey, N., & Sharkey, A. (2010a). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies, 11*(2), 161–190.

Sharkey, N. E., & Sharkey, A. J. C. (2010b). Living with robots: Ethical tradeoffs in eldercare. In Y. Wilks (Ed.), *Artificial Companions in Society: Scientific, economic and philosphical perspectives* (pp. 245–25). Amsterdam: John Benjamins.

Sharkey, N., van Wynsberghe, A., Robbins, S., and Hancock, E. (2017) Our sexual future with robots. A foundation for responsible robotics consultation report.

Sorell, T., & Draper, H. (2017). Second thoughts about privacy, Safety and deception. *Connection Science, 29*(3), 217–222. https://doi.org/10.1080/09540091.2017.1318826

Sparrow, R. (2002). The march of the robot dogs. *Ethics and Information Technology, 4,* 305–318.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Mind and Machine, 16,* 141–161.

Sung, J. Y., Guo, L., Grinter, R. E., & Christensen, H. I., et al. (2007). "My Roomba is Rambo": Intimate Home Appliances. In J. Krumm (Ed.), *UbiComp 2007, LNCS 4717* (pp. 145–162). Berlin: Springer.

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Science*, *194*(46), 17954–17958.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.

Turkle, S. (2017) Why these friendly robots can't be good friends to our kids. *Washington Post,* December 2017, Accessed December 2018 https://www.washingtonpost.com/outlook/why-these-friendly-robots-cant-be-good-friends-to-our-kids/2017/12/07/bce1eaea-d54f-11e7-b62d-d9345ced896d_story.html?utm_term=.12c983fe3db1.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (Eds.), *Advances in autonomous robotics systems* (pp. 85–96). Berlin: Springer.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The fight for a human future at the new future of power*. London: Profile Books Ltd.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1–59). New York: Academic Press.