Int J Soc Robot (2009) 1: 217–221
DOI 10.1007/s12369-009-0026-2

ORIGINAL PAPER

Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics

Mark Coeckelbergh

Accepted: 20 May 2009 / Published online: 5 June 2009

Abstract The development of pet robots, toy robots, and sex robots suggests a near-future scenario of habitual living with 'personal' robots. How should we evaluate their potential impact on the quality of our lives and existence?

In this paper, I argue for an approach to ethics of personal robots that advocates a methodological turn from robots to humans, from mind to interaction, from intelligent thinking to social-emotional being, from reality to appearance, from right to good, from external criteria to good internal to practice, and from theory to experience and imagination. First I outline what I take to be a common approach to roboethics, then I sketch the contours of an alternative methodology: ethics of personal robots as an ethics of appearance, human good, experience, and imagination.

The result is a sketch of an empirically informed anthropocentric ethics that aims at understanding and evaluating what robots do to humans as social and emotional beings in virtue of their appearance, in particular how they may contribute to human good and human flourishing. Starting from concrete experience and practice and being sufficiently sensitive to individual and cultural differences, this approach invites us to be attentive to how human good emerges in human–robot interaction and to imagine, possibilities of living with personal robots that help to constitute good human lives.

Keywords Personal robots · Ethics of robotics · Appearance · Human flourishing · Artificial intelligence

M. Coeckelbergh (✉)
University of Twente, P.O. Box 217, Enschede, The Netherlands
e-mail: m.coeckelbergh@utwente.nl

1 Introduction

There is an international trend in robotics from industrial applications towards robots that play a role in personal life. The development of pet robots, toy robots, and sex robots suggests a near-future scenario in which living with robots will be as habitual as living with TV, mobile phones, and internet. Such 'personal robots' will 'share physical and emotional spaces with the user' [1]. They could play a role in entertainment, education, household, and health care. Sometimes they are called 'social robots', and Turkle has proposed the term 'relational artefacts' [2]. Perhaps we will enter in relationships with robots [3]. In any case, personal robots are likely to have a significant impact on the quality of our lives and existence. How can and should we evaluate living with personal robots?

In this paper, I argue for an approach to ethics of personal robots that advocates a methodological turn from robots to humans, from mind to interaction, from intelligent thinking to social-emotional being, from reality to appearance, from right to good, from external criteria alone to good internal to practice, and from theory to experience and imagination. First I outline what I take to be a common approach to roboethics, then I sketch the contours of an alternative methodology: ethics of personal robots as an ethics of appearance, of the good life, and of experience and imagination.

2 A Common Approach to Roboethics

A common approach to roboethics (including ethics of personal robots) is characterised by the following methodological features: it focuses on the mind and reality of the robot and it understands roboethics as a branch of applied ethics

and ethics of the right. Let me explain these features and discuss the difficulties they incur.

2.1 Roboethics with a Focus on the Mind and Reality of the Robot

Unsurprisingly, many existing work in roboethics focus on the moral status and actions of the robot. For many moral philosophers, ethics is about holding someone responsible and about the rightness of one's actions, and then questions regarding moral status and action are central. We usually ascribe moral responsibility only to beings that have a sufficient degree of moral agency—whatever that means—and ask about the rightness of what that agent does, has done, or could do. Robots, then, present a challenge to our traditional theories of moral responsibility. Do they pass the moral agency test? Can they be held responsible [4, 5]? Should we grant them rights? Can we blame them, or even punish them? How should we treat them [6]? However, this focus on the morality of robots means that ethical questions concerning how humans interact with robots and how humans experience that interaction remain out of sight.

This neglect is reinforced by another development. Philosophers of robotics rightly feel that they need to engage with research in the field of artificial intelligence (AI), but this focus on AI usually implies that they base their ethical analysis on questions and assumptions that focus on the real and on the 'mind' of the robot. This is not so much the problem of AI; designers of artificially intelligent systems are often focused on what the system does, on what the programme needs to achieve. But *philosophers* of AI discuss problems of representation [7], rationality [8], or 'soul issues' [9]. Although they show the importance of appearance, the well-known 'Turing test' [10] and the 'Chinese Room' thought experiment [11] are *meant* to test how intelligent an artificial system is (not how intelligent it appears). Many researchers in roboethics, then, start from similar questions, since for them the moral status of the entity is at stake (see above). Is this robot really intelligent? Can a robot become conscious [12]? Could they count as moral agents [13]? If we know how the robot's 'mind' works [14], it seems, we can say more about roboethics. In this way, the emphasis is again on the robot and what the robot really is or 'thinks', instead on how robots appear to us, humans. Moreover, whereas there is growing attention for social and emotional aspects of robots, the social and emotional side of *human beings* and its significance for the ethics of human–robot interaction receives much less attention.

Apart from having less relevance, the usual approach to roboethics also faces a serious difficulty, since it relies on empirical *proof* of internal states. Consider the following example of the current, dominant approach to roboethics. In his recent paper, published elsewhere in this issue, David Levy [15] asks the question how we should treat robots. First, since he sees *consciousness* as being at the root of moral status, a major problem he has to deal with is how to 'detect' consciousness. In his paper he discusses several tests, ranging from the Turing test to the 'delay test' (based on delay conditioning; see work by Clark and Squire). In this approach, only when we can establish that robots are, or could be conscious, we can ask Levy's other question: should they have *rights*?

Note that apart from consciousness and the usual reference to the history of emancipation (first slaves, then women, are robots next?), he also provides an argument that is independent from moral status: if we hit a robot, we give the message to our children that this is how we should treat people. Although I think this argument does not only apply to robots but also to many other artefacts in many situations, I believe this is an interesting point, and one that is very much based on how we, as imaginative, sensitive, and appearance-driven beings, easily cross human/non-human borders in ethical life.

2.2 Roboethics as Applied Ethics and Ethics of the Right

Ethics is often understood as 'applied' ethics. In this view, if we are to evaluate human–robot interaction (perhaps including human–robot relationships), we should apply ethical criteria—moral principles provided by an ethical theory—to the problem or case at hand. This allows us to judge whether or not what goes on is ethically acceptable. In roboethics, this approach means that criteria external to what goes on in human–robot interaction are applied to it. For instance, if one thought that paying for sex is morally wrong, one would judge by this criterion that sex robots are morally not acceptable. Work such as the Roboethics Roadmap [16] typically takes an 'applied ethics' approach.

For designers, this 'external' approach implies that the aim is to try to 'build in' rules in their robot. Just as 'externalist'[1] moralists want *humans* to internalise the external moral rules, some moralist designers wish robots to have those rules 'in them'. For instance, taking their inspiration from Asimov's Three Laws of Robotics, they may want to build robots that avoid harm to humans, obey the orders of humans, and do not destroy themselves. However, such an approach to ethics—whether it is applied to humans or to robots—does not only run into trouble in concrete contexts (as Asimov's own stories show), but also neglects other kinds of ethical questions.

[1]When I use the term 'externalist' here, I do not refer to the usual meaning the term has in philosophical discourse, which is about motivation, reasons, justification, or the relation between mental states and the world. My usage of the term here is restricted to the relation between a particular practice or interaction, on the one hand, and moral norms, on the other hand.

In addition, ethics is often exclusively understood in terms of the right. Is this behaviour morally right or wrong? Thus, in roboethics it is asked if using military robots is morally right (and if the robot is morally responsible for what it does—see above), if it is morally right for a robot to harm a human being, if it is morally right to replace human sex workers by sex robots, etc. This approach leaves out broader ethical questions, such as which lives we want to live (with and without robots). It limits ethics to concerns about things that might go wrong in interactions with robots; it leaves out the question: what if all goes right, is it still good to live with these robots?

What is the alternative? Let me make some suggestions for *one* possible alternative approach.

3 An alternative: Appearance, Human Good, and Imagination

A possible alternative to the methodological orientation described above is to turn to an ethics of appearance and to an ethics of human good understood as emerging in experience and practice. Let me explain what I mean by these terms and explore the implications for ethics in general and roboethics in particular.

3.1 Roboethics as an Ethics of Appearance

Instead of asking how human-like robots can become in order to be able to ascribe agency, autonomy, and responsibility to them, I suggest that we start from studies of how humans interact with robots on the basis of apparent rather than real humanoid features (intelligence, consciousness, emotion, etc.). For example, does the face of a particular robot appear human, and if so, how do we experience this feature in interaction with that robot? What robots do to us, depends on how they appear to us, not on what is 'really' in their mind. For instance, existing robots are not sentient and lack feeling; nevertheless, when humans interact with some types of robots they may act and talk *as if* the robot has sensations and feelings. In a similar way, humans tend to attribute thoughts and beliefs to robots. These observations are relevant to the way we humans act and live. Rather than focusing our ethical worries on robots, let us worry about humans, about what *we* think, feel, and dream of. The approach to roboethics I propose is self-consciously anthropocentric instead of robocentric. Instead of a philosophy of mind concerning what robots really are or really (can) think, let us turn to a philosophy of interaction and take seriously the ethical significance of appearance. It is a turn from the 'inside' (what is 'in the mind' of robots) to the 'outside' (what robots do to us). Let us ask: What is the ethical implication of living with personal robots, that is, of interacting with them in a personal, social and emotional context? How do we perceive them, and what do they do to us as social and emotional beings?

To answer such questions, ethics of robotics can benefit from empirical studies of human–robot interaction, philosophical and psychological analysis of human emotional and social life, phenomenological studies of human–robot relationships, philosophy of technology, and emerging work in roboethics that starts paying more attention to appearance. For instance, we can learn from experiments that explore the minimal requirements for effective human–robot social interaction [17, 18], requirements that all depend on the appearance of the robot (e.g. certain facial features). We can benefit from discussions about Mori's 'Uncanny Valley' hypothesis [19], which again depend on appearance: the hypothesis is that robots who appear almost human,[2] can make us feel uncanny (Freud's term is *unheimlich*). We can also use research on interaction with humanoid robots and pet robots, such as work from Breazeal [20] and Turkle [2, 21]. We can learn from studies of perception and media [22] that even computers or simple objects can make us treat them human-like, from which we can conclude that human–robot 'social' interaction is less hard to achieve than those who focus on the 'mind' of the robot may think. We can also benefit from phenomenological contributions to ethics of human–robot interaction [23], a philosophical tradition which is particularly apt to talk about appearance. Of course, we should then not discuss the robot's consciousness, but how robots appear to our consciousness (or better: *my* consciousness, given the stress on the first-person perspective). And finally, we can benefit from the results of recent European research projects (e.g. ETHICBOTS) and networks (e.g. EURON), in so far as they teach us something about how we humans respond to robots.

Note that this approach is also much more in tune with what contemporary robot researchers and designers do: they think about what kind of interactions they want to achieve with their robot. They care less about consciousness, more about (inter)action and what this does to us. For example, someone who designs toy robots for children may reflect on how children will interact with the robot and what kind of interactions are appropriate for the child to engage in (e.g. at a certain age or stage of development).

Note that my approach, although directed towards the 'outside', is not at all behaviouristic. Let me explain this. First, my focus on appearance (Dutch: *verschijning*; German: *Erscheinung*) is meant to draw attention to the experience on the part of the human. The emphasis is not on what the robot does (its behaviour), but on what it does *to us*. This is not only a matter of observation, but also and

[2]Consider for instance Hiroshi Ishiguro's robots, who appear as 'copies' of himself and his daughter.

perhaps more of understanding, of understanding *humans*, without involving assumptions about what 'really' is 'in' the 'mind' of the robot. Second, therefore, my approach is not *anti*-behaviourist either. Behaviourism would mean that I claim that what we call mental states are really just publicly observable patterns of behaviour. I do neither agree nor disagree with that claim. Instead, I remain agnostic with regard to *any* theory of robot 'mind', since I hold that—at least when applied to the robot—such theories of mind are not very relevant to the ethical aspects of human–robot interaction.

3.2 Roboethics as an Ethics of Good and an Ethics of Experience and Imagination

Instead of limiting ethical evaluation of human–robot interaction to the question concerning the morally right, I propose that we also and especially consider the potential contribution personal robots could make to human good. Can human good appear in human–robot interaction (or relationships), or only in human–human interaction (and relationships)? Can human–robot interaction (relationships) contribute to human flourishing and happiness? Can such interactions constitute friendship, love, or relationships at all? Can they co-shape a flourishing community?

Let me explain this approach. What is 'human good' or 'human flourishing'? My use of these terms is inspired by neo-Aristotelian approaches to ethics, which focus on the question how we should live and what kind of moral habits and moral character we should develop rather than on the question how we should act (at a particular moment in time, in a specific situation). Moreover, I assume that some ways of life are better than others and that some goods are good for all humans (this is a so-called objectivist approach to what Aristotle called 'the good life'). Finally, I avoid the term 'the good life' often used in this tradition since I believe there are many ways of living that can be called good. Human good is plural. However, what are these human goods and what does this approach entail for evaluating personal robots?

The ethical questions asked above can be answered in at least two ways. One approach is to start from a certain conception of human good, of human flourishing, of happiness, of friendship, and of love. For instance, we may want to propose a list of criteria by which we are to judge the ethical aspects of personal robots. My own version of such a list is inspired by Martha Nussbaum's capability approach, and includes criteria such as health, imagination, affiliation, and play [24]. I have argued that these criteria can be used to evaluate robots and (other) artificially intelligent technologies such as assistive technologies [25]. This, in itself, is an innovation in ethics of technology since it provides a more precise and more 'workable' definition of human good than is usual given in ethical theory.

However, if such conceptions of good and such lists are taken as a priori criteria, pre-conceptions of good, then this approach is problematic. Above I rejected an 'externalist' approach to morality. Criteria might be so general or so remote from what goes on in human–robot interaction, that they are not very helpful. Therefore, we must take seriously the specificity of the new technology and of what it does to us by connecting the theory with practice in a stronger way. Rather than starting from the capabilities list as *a priori* moral norms, we need to start from concrete experiences and imagination of human–robot interaction and then discuss what good understood in terms of capabilities means. Existing research on human–robot interaction can be very helpful for this purpose, provided it is shares, or is interpreted from, the methodological perspective put forward in this paper. For instance, some interactions with personal robots may not only help people to develop their capability for play but also redefine that capability. Good is not independent of what happens in practice; it can only exist and flourish *in* practice.

In addition, we need an approach that is sufficiently sensitive to individual and cultural differences in experience and imagination. For instance, if at some point in the future young people will have been raised with personal robots (as they are now raised *with* personal computers, the internet, and mobile phones), they will not experience robots in the same way as elderly people who had and have to get used to them at a later point in their lives (as many have now difficulties with adapting to contemporary ICT technology and indeed life-with-such ICT). Furthermore, different individuals with different characters and personal histories may respond differently to the same robot, which as already been observed in nursing homes. And we observe already now that for instance Japanese people have a different attitude towards living with robots than people in Europe or the U.S. Robots.

Thus, whatever human good may be external to human experience, we must study, imagine, and shape concrete human living with robots as the locus where good may appear. Let us listen to people's experience and use our moral imagination to find out if there are possibilities of living with robots that enhance human flourishing and happiness. In this way, the design, use, and regulation of social robots can better contribute to good human lives.

4 Conclusion

From the discussion above I conclude that with regard to ethics of personal robots, a two-fold methodological shift is desirable and could usefully complement existing approaches. First, we should take seriously the ethical significance of appearance by not focusing exclusively on questions regarding the agency and 'mind' of the robot, but rather

on how we perceive robots and what they do to us as social and emotional beings. Second, instead of indulging in fantasies about moral robots with robot rights, we must be attentive to, and imagine, possibilities of living with personal robots that contribute to, and indeed co-constitute, good *human* lives in practice.

References

1. Cerqui D, Arras KO (2001) Human beings and robots: towards a symbiosis? In: Carrasquero J et al (eds) A 2000 people survey. Post-conference proceedings PISTA 03 (Politics and Information Systems: Technologies and Applications), pp 408–413
2. Turkle S (2005) Relational artefacts/children/elders: the complexities of cybercompanions. In: Android science workshop, Stresa (Italy). Cognitive Science Society, pp 62–73
3. Levy D (2007) Love and sex with robots: the evolution of human–robot relationships. Harper Collins, New York
4. Haselager WFG (2005) Robotics, philosophy and the problems of autonomy. Pragmat Cogn 13(3):515–532
5. Floridi L, Sanders JW (2004) On the morality of artificial agents. Minds Mach 14:349–379
6. Asaro P (2006) What should we want from a robot ethic? Int Rev Inf Ethics 6:10–16
7. Clark A, Grush R (1999) Towards a cognitive robotics. Adapt Behav 7(1):5–16
8. Clark A (2001) Reasons, robots and the extended mind. Mind Lang 16(2):121–145
9. Epstein RG (1999) Review of Hans Moravec, Robot: Mere machine to a transcendent mind. Ethics Inf Technol 1:227–236
10. Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460
11. Searl J (1980) Minds, brains and programs. Behav Brain Sci 3(3):417–457
12. Kitamura T, Tahara T, Asami K-I (2000) How can a robot have consciousness? Adv Robot 14(4):263–275
13. Torrance S (2007) Ethics and consciousness in artificial agents. Artif Intell Soc 22:495–521
14. Ishii K (2006) Cognitive robotics to understand human beings. Q Rev 20:11–32
15. Levy D (2008) The ethical treatment of artificially conscious robots. Paper presented at the 1st international conference on human–robot personal relationships, Maastricht University, June 13, 2008
16. Veruggio G (2006) EURON roboethics roadmap (release 1.1). EURON Roboethics Atelier, Genua
17. Bruce A, Nourbakhsh I, Simmons R (2002) The role of expressiveness and attention in human–robot interaction. In: Proceedings of the 2002 IEEE international conference on robotics & automation, Washington, DC, May 2002, pp 4138–4142
18. Duffy BR (2003) Anthropomorphism and the social robot. Robot Auton Syst 42:177–190
19. Mori M (1970) Bukimi no tani (The uncanny valley). Energy 7(4):33–35. (Original in Japanese, translated by MacDorman KF & Minato T)
20. Breazeal C (2003) Emotion and sociable humanoid robots. Int J Human–Comp Stud 59:119–155
21. Taggart W, Turkle S, Kidd CD (2005) An interactive robot in a nursing home: preliminary remarks. In: Android science workshop, Stresa (Italy). Cognitive Science Society, pp 56–61
22. Reeves B, Nass C (1996) The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, Cambridge
23. Ramey CH (2005) "For the sake of others": the personal ethics of human–android interaction. In: Android science workshop, Stresa (Italy). Cognitive Science Society, pp 137–148
24. Nussbaum MC (2006) Frontiers of justice. Harvard University Press, Cambridge
25. Coeckelbergh M (2009) Health care, capabilities, and AI assistive technologies. Ethic Theory Moral Pract (forthcoming)

Mark Coeckelbergh is Assistant Professor at the Philosophy Department of the University of Twente, The Netherlands. He is also Senior Researcher with the 3TU. Centre for Ethics and Technology. Previously he has taught philosophy at Maastricht University, The Netherlands. His publications include Imagination and Principles (2007), The Metaphysics of Autonomy (2004), Liberation and Passion (2002), and articles in the areas of ethics of technology, engineering ethics, bioethics, and political philosophy. Currently he has a particular research interest in ethics of robotics and supervises a Ph.D. project on 'Carebots and the Good Life'. In 2007 he received (with Jessica Mesman) the NVBe prize from the Dutch Association for Bioethics.