Studies in Ethics, Law, and Technology

Volume 4, *Issue* 3 2010 *Article* 2

THE CONVERGENCE OF THE PHYSICAL, MENTAL AND VIRTUAL

Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations

Mark Coeckelbergh, *University of Twente*

Recommended Citation:
Coeckelbergh, Mark (2010) "Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations," *Studies in Ethics, Law, and Technology*: Vol. 4 : Iss. 3, Article 2.
Available at: http://www.bepress.com/selt/vol4/iss3/art2
DOI: 10.2202/1941-6008.1126

Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations

Mark Coeckelbergh

Abstract

Under what conditions can robots become companions and what are the ethical issues that might arise in human-robot companionship relations? I argue that the possibility and future of robots as companions depends (among other things) on the robot's capacity to be a recipient of human empathy, and that one necessary condition for this to happen is that the robot mirrors human vulnerabilities. For the purpose of these arguments, I make a distinction between empathy-as-cognition and empathy-as-feeling, connecting the latter to the moral sentiment tradition and its concept of "fellow feeling." Furthermore, I sympathise with the intuition that vulnerability mirroring raises the ethical issue of deception. However, given the importance of appearance in social relations, problems with the concept of deception, and contemporary technologies that question the artificial-natural distinction, we cannot easily justify the underlying assumptions of the deception objection. If we want to hold on to them, we need convincing answers to these problems.

KEYWORDS: robots, artificial companions, ethics, empathy, vulnerability

Author Notes: I would like to thank the organisers and participants of the 2009 "ICT That Makes The Difference" conference in Brussels, the anonymous reviewers, John Stewart (Compiègne University of Technology), and the people who commented on a previous version of this paper presented at the January 2010 COGS seminar at Sussex University (Steve Torrance, Tom Froese, Blay Whitby, Margeret Boden, and others) for their advice and suggestions for improvement. My visit to Sussex was kindly sponsored by the EUCogII network.

Introduction

Suppose that we started to see some robots not as mere objects, artificial slaves, or instruments for human purposes, but that instead they appeared as more-than-things, as companions or perhaps as friends or even partners. Imagine, furthermore, that we also treated them as such, that is, in a very similar way as we now treat our pets, friends or partners. We would talk with robots, worry about them, be angry at them, play with them, live with them. Even sex and love would be possible.

This scenario raises many questions. For example, the idea of sex with robots can be seen as morally repugnant[1], and offensive even to think or write about it. And, what does 'love' mean when someone talks about love with robots? Surely this is not *real* love, which can only exist between humans? In this paper, I limit my discussion to the issue of artificial companionship. Are these robots *real* companions? What kind of companionship is going on here? It seems that those who were to live with these robots would be deceived: they would believe that they interact with a real pet or maybe a real human. Is this deception, and if so, is this kind of deception (or self-deception) ethically acceptable?

Whatever we think of it, the idea of humans engaging in companionship relations with robots is not science-fiction. Robots play a role in many domains (Veruggio, 2006), including personal and social contexts and practices such as entertainment, sex, nursing, and education. Whereas companionship with highly human-like humanoid robots may be mere speculation or at least something that belongs to the distant future (the development of highly human-like humanoid robots is still in its infancy), it is likely that some of us will live with robot pets and other robotic 'companions' in the near future – at least, if current trends and technological developments continue. They may not look exactly like humans and may lack many human capacities, but they are meant to function as companions. Before I question and discuss the use of this term as applied to robots, let me say more on the state-of-the-art in the field of 'companion robotics'.

Robotic companions

People already 'keep' commercially available robots in their homes (e.g. the robot Pleo: a robot dinosaur with the behaviour of a week-old baby). There have been experiments with robots in care for the elderly (see Wada et al, 2006 on the therapeutic baby seal robot Paro; see also Bickmore et al, 2005), and they have been used in autism therapy (Robins et al, 2006). Also, given the history of using artificial devices and information technology for sexual purposes, there is a future

[1] Consider the controversy stirred up by Levy's book on sex with robots (Levy, 2007).

for interactive sex robots (Levy, 2007). More generally, use of robot companions can be expected to increase in the near future, if robot companions will be more (artificially) intelligent and capable of interacting with humans in a more meaningful way. Progress has been made in creating embodied 'relational agents' that use face-to-face conversation (Bickmore et al, 2005) and 'sociable' robots have been designed that appear to have emotions, such as the robot Kismet, with its facial expressions (Breazeal, 2003), and the successors Leonardo and Huggable developed by Breazeal and others at MIT. For example, Huggable is designed to be capable of 'active relational and affective touch-based interaction with a person' (Stiehl et al, 2006). Some of these robots can *learn* by interacting with humans. Often learning robots are child-like humanoid robots, such as KASPAR (designed by Dautenhahn and others at the University of Hertfordshire), CB2 (a child robot that develops its social skills by interacting with humans, designed at Osaka University in Japan), and iCub (a robot with the size of a 3.5 year old child that is inspired by theory of embodied cognition and developed by a European consortium). Moreover, such 'social robots' may not only act as artificial pets or even 'partners' (Breazeal, 2003), but also contribute to information services, home security, medicine and health care, elderly care[2], and household tasks (Floridi, 2008; Dautenhahn et al, 2005). For instance, they could guard the house when the owner is on holiday, assist nurses in hospitals, monitor people with a precarious medical condition, and help elderly people, allowing them to live 'independently' in their own environment.

Concrete steps are being taken to develop robots for these purposes. Apart from the ongoing development of commercial sex robots (for example the recently presented sex robot Roxxxy, which, according to inventor Douglas Hines, is also able to talk about football), the areas of health care and elderly care receive much attention from research and funding institutions. I already mentioned the Huggable developed at MIT, which can be used for therapeutic purposes. In the EU there has been the Cogniron project (Cognitive Robot Companion, 6th Framework Programme) and more recently the Companionable project (Integrated Cognitive Assistive & Domotic Companion Robotic Systems for Ability & Security), funded by the EU 7th Framework Programme. Aiming to enhance the quality of life of elderly and disabled persons, and help them to live independently in their own home, the robotic system would monitor, recognise emotional states of persons, distinguish between normal and exceptional behaviour, remind persons to take their medication, and so on.

In response to these technological developments and future scenarios, multi-disciplinary work on human-robot relations and social robotics has grown significantly (e.g. Dautenhahn, 2005, 2007; Breazeal, 2003, Duffy, 2003). This

[2] The use of robots in elderly care has been proposed as one response to population ageing in Western countries and Japan.

DOI: 10.2202/1941-6008.1126

paper contributes to work in that area by providing a conceptual analysis of companionship and by discussing the ethical problem of deception in relation to this issue. Let me first elucidate my research question.

Questions

In my introduction, I suggested some questions concerning artificial companionship. Is it appropriate to say that a robot is a 'companion'? Can it be a companion at all? On what does the possibility of artificial companionship depend? We can think of many conditions, such as the capacity for a certain level of interaction and communication or human-friendly and safe behaviour. The robot would also need to have the opportunity to exercise these capacities in daily human, social environments (as opposed to laboratories or computer simulations). In this paper, I explore my intuition that one of the necessary (but not sufficient) conditions for companionship concerns empathy.

More precisely, my questions are: Can robots become companions, and if so, on what conditions and which are the ethical issues that might arise in human-robot companionship relations? I will argue that the future of robots as companions depends (among other things) on the robot's capacity to be a recipient of human empathy and that one necessary condition for this to happen is that the robots mirrors our own, human vulnerabilities. Moreover, I will show that we might have the intuition that this raises the ethical issue of deception. But, given the importance of appearance in social relations, problems with the concept of deception, and contemporary technologies that question the artificial-natural distinction, we cannot easily justify the underlying assumptions of the deception objection.

Note that I limit the scope of my argument to *one* condition for companionship and *one* condition for empathy. Human beings, companionship, and empathy are much more complex than that, but here I will pursue one dimension of companionship specifically; there may be many more criteria for companionship, as there are more aspects to empathy.

Empathy in human-robot relations

Can *robots* have the capacity for *empathy*? If I suggest that empathy may be a condition for companionship, I need to answer at least two questions: (1) *who* (or what) exercises empathy towards *whom* (or what) and (2) *what* do I mean by the term 'empathy'?

Robot empathy and human empathy

Who or what has a capacity for empathy? In human relations, empathy can go two ways: person A can exercise empathy towards person B and vice versa. But what about robots? Some might say that in the future humanoid robots will be capable of empathy. I doubt it. But whether or not it is possible or will happen, as it stands there are good reasons why robot empathy should not be taken as a condition for companionship. First, to require that robots have the capacity for empathy would be an 'unfair' requirement since we do not require it from all human companions. Very young children and some adults with a mental disorder lack empathy but we still treat them as companions, not as 'mere things' or instruments for our purposes. Secondly, although pets like dogs or cats might be capable of *some* empathy, we do not demand it as a condition for companionship: it suffices that we humans exercise our empathy towards them. Companionship relations, in contrast perhaps to friendship and love relations, need not be entirely symmetrical in this sense: one-way empathy is enough as one necessary condition to sustain the companionship relation. Therefore, I propose that we agree on the minimal requirement that robots can be *recipients* of human empathy as a necessary condition for human-robot companionship.

Understood in this way, it seems that the empathy condition is already met. People who interact with the robots mentioned in the introduction already 'put themselves in the robot's shoes'. Very much like pet owners, they take into consideration what the robot 'thinks', 'feels', or 'wants'. But what does this mean? If people empathise with robots, what kind of empathy is this? Is this 'real' empathy? What is empathy?

Empathy as cognition v. empathy as feeling

Inspired by the discussion on the nature of emotions[3], let me distinguish between two approaches to, and definitions of empathy. The first approach views empathy as a cognitive matter. The question is: What is in the mind of the other? We want to achieve knowledge about the (content of) the mind of the other. This kind of empathy is less applicable to relations with robots, since (1) generally we do not suppose that they have a 'mind' and (2) *if* they had one, then how could we know it? How could we know – to employ Nagel's phrase – *what it is like to be a* robot and what it is like to be *that particular robot*? There are limits to our (cognitive) imagination.

[3] In theory of emotions there are (among other tensions) two opposing strands. According to cognitivists, emotions are cognitive: they are attitudes, beliefs or judgements (de Sousa, 1987; Solomon, 1980; Nussbaum, 2001; Goldie, 2000), whereas feeling theories understand emotions as awareness of bodily changes (James, 1884; Prinz, 2004).

DOI: 10.2202/1941-6008.1126

Note that the sceptic's objection also casts doubt on the possibility for this kind of empathy to emerge in human-human and human-animal relations: how can we know what it is to like a particular animal or a particular other person? Consider the famous philosophical problem of 'other minds'. Can we be so sure that others have a mind very much like ourselves? Can we have access to the experience and consciousness of others? Do they have an inner life?

Thus, if we defined empathy in a way that renders it a cognitive, epistemological, or mind problem and if we had no solution to the epistemological problem, robots would not be able to be recipients of human empathy.

However, another way of defining empathy: as an imaginative-emotional function. We can always try to *imagine* the mind of other humans or of robots and this empathy is not so much a matter of (certain) belief or knowledge concerning mental states but of feeling: we imagine how the other (human or robot) *feels*. The 'content' that counts here is not what is 'in the mind' of the robot, but what humans feel when they interact with the robot. Here there are no large epistemological problems. Our emotional experience understood in terms of feeling is – by definition – accessible to ourselves. If it was inaccessible, we would not feel it. Therefore, if we define empathy in terms of empathy-as-feeling instead of empathy-as-cognition, it becomes possible that robots function as recipients of our empathy.

But what kind of feeling is it? It is felt that the robot is 'one of us'. Human 'fellow feeling' imagines robots, animals, and humans as 'fellows' rather than as alien minds that need to be known. The 'other' becomes less 'other'; in the centre of this mental operation is what is shared rather than what is different. Moreover, not the individual mind but the social aspect is stressed. There may not be a large similarity in terms of particular properties (for example, the robot has an artificial 'body' and we have a biological body), but the robot is seen as a social fellow, an entity that belongs to our social world. It *appears* as one of us.

I borrow the term 'fellow feeling' from the 'moral sentiment' tradition, which stresses how feelings shape the social and moral life. Hume (2000 [1739/40]) and Smith (1976 [1790]) argued that rather than rational choice, our conduct towards others is a matter of feelings for others. The emphasis in this tradition is on the problem of how the social life emerges from what Hume called 'sympathy'. Sympathy for others – which we can 'translate' in terms of empathy-as-feeling – renders it possible that we can live together. This approach can inspire to a more social understanding of ethical problems in human-robot relations: we should be less concerned about what the robot is (rational arguments assessing the moral status of 'the robot') and more about how we want to live together, given that we *already* engage in particular social relations with them in particular contexts. Instead of regarding the robot from a point of 'Nowhere' (as

Nagel put it, see Nagel, 1986), we can try to understand what goes on starting from the feelings of humans who interact with robots and build up social relations with them. From this perspective, one important and necessary condition for robot companionship has to do with whether or not humans meaningfully regard the robot as a 'fellow', whether or not they *feel* that the robot is a member of their social world. This approach takes seriously human experience in and of human-robot relations. It does not *a priori* dismiss such relations as being a matter of deception (see my argument below).

Note that perhaps there is no hard distinction between empathy-as-cognition and empathy-as-feeling. In contemporary cognitive science and theory of emotions, efforts are made to combine both approaches: if cognition is embodied and to feel cannot do without the awareness and evaluation of the feelings felt (or if emotions have a cognitive *and* feeling component), then the distinction I started from may well turn out to be too strict. More work is needed to spell out the relations between cognition, emotions, and feeling. It might be helpful to distinguish between different 'levels' of empathy. For example, Wallach and Allen suggest that empathy involves 'lower level' mirroring skills, emotions, and 'higher-level' imagination (Wallach and Allen, 2008: 164). I also acknowledge that it might be interesting to integrate more insights from cognitive science (embodied approaches) and theory of emotions into these arguments. I shall explore this further elsewhere. However, even if we could fashion a more adequate conception of empathy in terms of the relation between cognition and feeling, such a conception would not be incompatible with an account of empathy as fellow feeling. Whatever the exact nature of the relation between feeling and the rest of our mind-body make-up, an emphasis on fellow feeling allows us to move away from a philosophy of robot mind (consider also the ugly phrase 'Theory of Mind' used in that kind of discussions rather than empathy) towards a more *social* theory of human-robot relations that pays more attention to feeling – whatever its relation to cognition may be.

Vulnerability mirroring

Let me say more about empathy as 'fellow feeling' from a hermeneutic-phenomenological perspective rather than from a (cognitive) science perspective angle. When do we regard someone or something as a 'fellow'? What creates this fellow feeling? On what does it depend? I suggest that a necessary condition for this to happen is the requirement that the robot (or, for that matter, the animal or the human) mirrors our own, human vulnerability. Let me explain what I mean.

Our embodied existence renders us vulnerable beings. Human empathy is partly based on the salient mutual recognition of that vulnerability: this is what we (among other things) share as humans; this is what makes you 'like me' or 'one of

DOI: 10.2202/1941-6008.1126

us'. In this sense, we are each other's 'vulnerability mirrors'. We can feel empathic towards the other because we know that we are similar as vulnerable beings. If we met an invulnerable god or machine that was entirely alien to us, we could not put ourselves in its shoes by any stretch of our imagination and feeling (see also below).

Note that this 'vulnerability mirroring' can also happen in relations with a fictional character. Nussbaum has suggested in *Poetic Justice* (1995) that by reading literature we come to understand similarities between ourselves and the other and realise that we share vulnerability. More generally, empathy has to do with recognising all kinds of similarities, not just shared vulnerability or related existential features such as mortality. Nussbaum has argued that empathy allows us 'to comprehend the motives and choices of people different from ourselves, seeing them not as forbiddingly alien and other, but as sharing many problems and possibilities with us' (Nussbaum, 1997: 85). Thus, vulnerability mirroring is part of a more general mirroring process which establishes an empathic connection between humans, and which is also further stimulating that empathic process. The result is that we come to see the other as a fellow, as 'one of us' rather than an alien other. Consider also the fact that we feel most empathic towards those who are 'nearest' to us – not so much geographically but in terms of them having similar kinds of lives, problems, and aspirations.

However, vulnerability mirroring is not strictly limited to human-human relations. Our social world, in so far as it is constructed by empathy as fellow feeling, does not coincide with an all-human sphere. It helps vulnerability mirroring if there is already a relation, perhaps based on shared properties. As Hume wrote about what he calls 'sympathy' in *A Treatise of Human Nature*: 'The stronger the relation is betwixt ourselves and any object, the more easily does the Imagination make the transition […]' (Hume 2000 [1739/40], Book II, Section XI).

However, we should not conceive of similarities in terms of properties alone but also consider the active, practical side. The etymology of 'companion' links the word to 'eating the same bread': it refers to shared needs in addition to shared practices of fulfilling these needs. Some pet animals such as dogs and cats also mirror our vulnerability: they have their weaknesses, their problems, their characters, their little sufferings, their needs, etcetera. This is a necessary condition for us to accept them as companions. They are the mirror in which we understand that we (humans) are vulnerable and it is because we see their (animal) vulnerability as related to our own that we come to see them as fellows.

To the extent that robots can function as 'vulnerability mirrors', they fulfil at least one of the necessary conditions for companionship. *If* they meet this criterion, they can be considered as 'fellows' rather than 'mere things'. Of course many current robots do not meet this criterion because the differences between the

robot's vulnerability and the vulnerability of the human are too large. For example, we may feel that we have little in common with a vacuum cleaner robot (e.g. the Roomba). But many pet robots that are used today meet the empathy as vulnerability mirroring criterion: they are regarded by the humans who interact with them as fellows and receive human empathy, and this is partly and necessarily due to the 'vulnerability mirror' operation of empathy-as-feeling: humans do not so much *know* but *feel* a shared vulnerability. Provided that robots fulfil other necessary conditions (such as human-friendly behaviour) they will be regarded as companions.

Note that empathy is not only about suffering; we can also empathise with joyful others, winning others (consider people empathising with their winning football team), and so on. However, this is only possible if our friends or heroes are still vulnerable like us to a sufficiently high degree. Humans (including football players and other celebrities) meet this criterion effortlessly. And we can sympathise with a human-like suffering god, but not with a perfect 'philosopher's god' who would be incapable of (human-like) suffering or joy. Robotic vulnerability mirroring, then, can only be 'successful' in terms of establishing artificial companionship if it strikes that particularly human balance. Not only would these robots have to be qualitatively vulnerable in a way that resembles human vulnerability (for example have a 'body' that is vulnerable in a similar way biological bodies are), the designers would also have to get the degree of vulnerability right. If robots are too vulnerable and (as arguably most current robots are, dependent as they are on human control and energy supply), empathy cannot find its object since there is too much dissimilarity; but if they are not vulnerable enough, they cannot receive our empathy either. If robots were titans of steel and silicon or entirely virtual robots that enjoy immortality in a sphere of invulnerability, they would be far too invulnerable and -non-tragic to deserve our human empathy. Whatever else they may be or do (they might be benevolent or of the 'Terminator' type and they may be useful to us or not), we would not call them companions.

To conclude, I have suggested that the emergence of companionship relations between humans and robots depends on empathy. Exploring this intuition, I have argued that this empathy must be understood in terms of fellow feeling (experienced by the human) based on what I call 'vulnerability mirroring'. In practice and given current levels of technological development, this implies that robotic companions would need to be designed to mimic biological entities like small children or pets, which we see happening right now. In the future, vulnerability mirroring might be achieved with more intelligent and more human-like artificial entities.

Note that so far I did not make a normative, programmatic claim. The development of robotic companions may well be unacceptable for various

DOI: 10.2202/1941-6008.1126

reasons. My intention was only to outline and discuss one condition for robotic companionship. However, now I will discuss an ethical objection-deception.

The deception objection

We might object that there is an ethical problem with this kind of empathic relation: since robot pets *imitate* their biological cousins, they seem to *deceive* humans. They make us believe that they deserve our empathy, but they are 'mere machines'. The human vulnerability might be real, but *they* do not really mirror that vulnerability since they are not living beings. The mirror deceives and deception is morally unacceptable.

Before I continue to discuss this objection, let me note that of course there are other moral problems with human-robot relations and robot companionship, and there are good reasons to question the development of intelligent autonomous robot companions.[4] But here, I will focus on the problem of deception.

This particular objection against vulnerability mirroring must be related to more general concerns about deception by robot companions. Consider Sparrow and Sparrow's objections against using robots as companions in elderly care: they argue that when elderly people mistakenly believe that the robot really cares about them, is pleased to see them, and so on, this delusion is morally problematic: it is 'deception' and this deception is morally 'bad' for mainly the following reason:

> What most of us want out of life is to be loved and cared for, and to have friends and companions, not merely to *believe* that we are loved and cared for, and to *believe* that we have friends and companions, when in fact these beliefs are false. (Sparrow and Sparrow, 2006: 155).

In other words, robots may provide elderly care services but they *don't care*: they don't really care *about* the person they 'service'. I agree that this is the case (the robots are not conscious, how *could* they have genuine concern for anything at all?) and I share the intuition of the authors concerning deception and its morally problematic character. However, if the objection were to be held against my vulnerability argument, there are some major problems once we try to justify the objection. Let me offer several possible responses to the deception objection *as related to the vulnerability argument*. (And as a by-product, these responses are also relevant to the general deception objection.) They concern (1) the importance of appearance for social life, (2) conceptual difficulties with the term deception,

[4] Consider for instance the problem of responsibility given that many of these companion robots are or will be autonomous systems that are to a large extent not under direct human control. (Compare with similar responsibility problems in military robotics).

and (3) problems with the distinction between, and evaluation of, artificial and biological vulnerability.

Appearance and the social life

First, it is controversial to claim that lying is always morally unacceptable. Many of us would agree that it is permitted to lie in the well-known inquiring murderer case[5] or in cases that involve serious threats to the life of (at least) one's partner, children, friends, or oneself. It is also highly doubtful that making-believe is always morally problematic. In fact, we humans do it a lot. We do it not only in some particular practices such as poker or other games; to some extent 'deception' is present in all social interactions when we take on social *roles* which, by definition, involve role playing. In those interactions the difference between 'reality' and 'appearance' or between 'inauthentic' and 'authentic' is not so clear. Appearance is the glue of social life and involves learned, often involuntary[6], salient 'deception'. There is no perfect and unproblematic match between the social roles we play and the person we 'really' are. One may regret this and conclude from society's 'corruption' that we should retreat from society[7], but this can hardly be a solution for all of us – nor for the robot, which is supposed to participate in social life. Moreover, part of social learning is imitation and it is not *obvious* why imitation of social behaviour by a robot is more problematic than (learning by) imitation by humans, for instance by children. And if there is 'deception' on the part of the human, then what exactly is the moral difference between giving illusions to a young child (for instance by means of stories about imaginary beings or by means of giving them stuffed animals) and giving social-robotic illusions to an elderly person with mental capacities that reach a similar level as the young child's? Does it matter for deception if the robot looks human or not?[8] Does the problem change if it is a *virtual* robot or a real robot? I do not know the precise answers to these questions but at least those who use the deception objection should provide them and make finer distinctions

[5] Someone knocks at your door and asks where the person is he wants to kill. You know where the person is and you know that the man intends to kill that person and will probably succeed in doing so. Do you tell the truth about where the person is? The case is often used to discuss Kant's view that lying is never permissible.

[6] It is a fact of social life that we neither fully control our appearance nor the social consequences of that appearance.

[7] Sometimes this view is attributed to Rousseau, but his view is more complex: society corrupts us in various ways but civil society (the political life as citizens) can liberate us.

[8] Some worries may be avoided by building robots that look like pets rather than humans. The 'uncanny valley' hypothesis (Mori, 1970) suggests that we fear robots that look nearly human more than dolls or humans (or robots that would not allow us to differentiate between them and humans).

DOI: 10.2202/1941-6008.1126

between different kinds of robots and practices. I feel that it would be unfair to demand of (all) companion robots that they are more 'authentic' than humans – adults and children – in particular situations, contexts, and stages of life. In the case of relations between robots and elderly people with cognitive limitations, for instance, we need to think harder about the moral quality of relations between that-kind-of-objects (appearing as more than things) and that-kind-of-subjects (appearing as not enjoying full subjectivity).

Deception?

Secondly, it is not obvious that there is 'deception' in the case of vulnerability mirroring and the emergence of fellow feeling. Deception hides truths, and truth is a criterion for knowledge. But the kind of empathy at work here is a matter of feeling and it is not obvious that feelings can be 'true' or not. One would need to specify a criterion outside the empathic process that would allow us to evaluate the 'truth' of the feeling. One criterion could be 'the truth about the robot'. People who entertain the objection mean that the robot is 'just a machine'. But precisely this 'observation' can be questioned given people's experience of some robots as being more-than-machines. 'The truth about the robot' is not something that exists entirely independent of human subjective experience and cannot be limited to the scientific view but must include user experience. Another external criterion might be the requirement that the feeling originates from a 'true' or 'authentic' self. But what is this authentic self? This has been a long-standing philosophical-psychological issue; the meaning of the term cannot be taken for granted.[9] We should be reminded that (human) others are not always transparent to us and neither are we always transparent to ourselves. Moreover, it is questionable if the self can be defined absolutely independently from the social world – which in this case might include a robot.

One might further object that here persons are not deceived by the robot but *by themselves* and that this is the morally problematic point. However, the notion of self-deception is notoriously controversial, since it seems to imply that at a given time it is true that a person P knows the truth (about the properties of the robot) and does not know the truth (believes that the robot really cares etc.). But even if this is psychologically possible (as opposed to logically), then surely this defuses the objection since it implies that the person has the possibility to gain access to the truth *if (s)he really wanted to*. In that case, it seems, there is no hard *moral* problem concerning the human-robot relation.

[9] The literature on authenticity and self is vast, including work by Rousseau, Kierkegaard, Marx, Heidegger, Sartre, Fromm, and numerous contemporary discussions (for example Trilling, 1972; Taylor, 1992; Anton, 2001).

In fact, in human-robot relations where the humans are adults with no mental disabilities, these humans *know* that the robot is a machine, a non-human and at the same time they *feel* that the robot is more than that – and know that they feel in this way – and they do not have any psychological or moral problem with this. Being highly social animals, we are used to live by appearance and we can cope with different levels of experience.

In this respect, it is telling that when research shows that socially communicative robots are more likely to be accepted as companions, it is their *perceived* sociability that makes users accept them (Heerink et al, 2006). Reeves and Nass (1996) showed us that users even treat ordinary computers in a social way, like real people, but depending on appearance. The social importance of appearance is also confirmed by experiments by Robins, Dautenhahn and others in which autistic children prefer *not* to interact with a robot that appears human (Robins et al, 2006). They prefer 'mechanical' non-social interaction. So, it seems that if we do *not* relate to robots on the basis of the appearance of sociality, we have a (social) problem. But robots are also studied as interactive 'new media' (see for example Shinozawa et al, 2003) which has foregrounded cultural differences in how people respond to robot appearances.

I concede that in the case of humans with significant mental disabilities (elderly or not) – i.e. people with very limited cognitive capacities – there may be a moral problem. However, as I said above, I do not have a straightforward answer to the question why giving robotic companions or other 'more-than-objects' (e.g. dolls) to these impaired humans is more morally problematic than giving the same robots or dolls to young children with similar mental abilities, provided, of course, that their emotional and other well-being is served and enhanced and that these robotic companions do not *replace* human contact and care.[10] (Compare with our treatment of young children: when we give them dolls, the dolls are not meant to replace human contact and care).

Thus, if one wishes to defend the value of truth – and perhaps philosophers have a moral duty to do so – then at least one has to take into account how truth is treated by humans in existing social, technological, care practices. Otherwise, philosophy risks to be irrelevant to concrete moral experience and practice.

[10] On this point concerning replacement I agree with Sparrow and Sparrow (2006).

DOI: 10.2202/1941-6008.1126

Biological and artificial vulnerability

Thirdly, the deception objection to the *vulnerability* mirroring argument involves at least the following problematic assumptions:

1. only biological vulnerability is real
2. biological vulnerability is more real than artificial vulnerability
3. biological entities are more vulnerable than artificial ones
4. biological life is more valuable than artificial 'life'

Can these assumptions be justified?

Firstly, there is no doubt that machines are also vulnerable. Whereas cyberspace and information technologies may well *aim* at invulnerability (and perhaps immortality, as in transhumanist versions of techno-utopian worlds), they depend on software and hardware that are very vulnerable indeed. Hardware can easily be destroyed and computer programmes can 'crash' or get damaged in various ways. And cyberspace, which might seem almost invulnerable considered as an immaterial sphere disconnected from earth, is very dependent on hardware and software all over the world – and on the humans who use, design, maintain and monitor the supporting systems. Robots are as vulnerable as their hardware and software is (if this distinction is still meaningful at all)[11]. Moreover, there are at least metaphorical parallels to biological vulnerability. Software risks can be described in epidemiological terms (viruses that spread etc.) and hardware is as material as the human body. Why do we hold on to the intuition that biological 'hardware' is more valuable and that its vulnerability is more 'real'? If these technological vulnerabilities are so significant for the way we work and the way we live, how 'unreal' or unimportant are they? In any case those who make these assumptions must provide an answer to these questions.

Secondly, the very border between the 'natural' and the 'artificial' is continuously called into question by technological developments in medicine, biology, and converging fields: intelligent prosthetic devices, synthetic biology, nano-pills, or brain-computer interfaces make us into 'cyborgs' to some extent: mixtures of biological and technological substances. Thus, in the future our intuitions concerning the border between natural life and artificial 'life', , and the importance of that distinction might change. Those who argue against robot companionship by relying on current values should at least take into account the very possibility of moral-technological change: technology is not neutral towards morality but often changes our morality. Consider what the contraceptive pill has done for our evaluation of sex before marriage or what social network sites do to

[11] Recent developments in robotics (e.g. when 'bio' and 'techno' merges) and cognitive science (embodied cognition) call into question the software/hardware distinction.

the value of privacy. While it is difficult to predict how future robots will change our lives and our values, one should not assume a static view of morality.

Conclusion

I conclude that the future of companion robots depends – among other criteria – on the robot's ability to invite the development and exercise of human empathy towards the robot, in particular on empathy-as-feeling which is conditional upon vulnerability mirroring.

Most current robots do not meet this criterion and still appear to us as 'mere machines', but this is changing. Some 'owners' of robots come to see themselves increasingly as 'keepers' of the robot, as its 'friend', or even as its 'partner'. People involved in human-robot relations increasingly replace the neuter pronoun 'it' by the personal pronouns 'he' and 'she'. (In the English language a similar change has happened with regard to babies: today most people prefer to use personal gendered pronouns). Part of what goes on here is the development of (one way) empathy: the robot is considered as an other, it is felt that the robot is part of our social world, and this invites humans to put themselves in the robot's shoes and evaluate and adapt their behaviour accordingly. "What would the robot think if I did this?" "The robot would feel unhappy if I didn't do this." "I feel guilty neglecting my robot's wishes." Are these changes to be welcomed?

I sympathise with the response that there is something morally problematic going on here. It seems that people who interact with the robot mistake it for a human person and at first sight it seems that this mistake is a moral problem. However, I have shown that while the deception objection to this kind of empathy is intuitively appealing, it is hard to justify given how we treat appearance and truth in our social practices, given conceptual problems related to deception, and given contemporary developments in information technologies that show how *real* technological vulnerabilities can be (and how important they are for our lives), and suggest that our natural/artificial distinction might be more 'artificial' than we think it is. Therefore, if we want to hold on to our intuitions about truth and deception with regard to robotic companions, we need a better justification that answers these objections.

If we still feel that deception is bad and that companion robots deceive humans, we should show how and why this is the case. We should argue which robots in which human-robot relations in which practices and contexts are doing something that counts as deception, then compare what happens in other relations and contexts, and provide a better justification of why (that kind of) deception is bad. These arguments should be informed by a sound understanding of existing social-emotional practices and the role of appearances and 'more-than-objects' in

DOI: 10.2202/1941-6008.1126

these practices. In the meantime, some of us will treat their pet robots as companions and grant them their empathy – without worrying about the truth of their (human) feelings or the artificial nature of their favourite vulnerability mirror.

References:

Anton, C. (2001). *Selfhood and Authenticity*. Albany, NY: State University of New York Press.

Bickmore, T. W., Caruso, L., Clough-Gorr, K., and Heeren, T. (2005). It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers* 17(6): 711-735.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems* 42: 167-175.

Dautenhahn, K., et al. (2005). What is a Robot Companion - Friend, Assistant, or Butler? *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.* (IROS, 2005): 1192-1197.

Dautenhahn, K. (2007). Methodology and Themes of Human-Robot Interaction: A Growing Research Field. *International Journal of Advanced Robotic Systems* 4(1): 103-108.

de Sousa, R. (1987). *The Rationality of Emotion*. Cambridge, MA: MIT Press.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42 (3-4): 177-190.

Floridi, L. (2008). Artificial Intelligences's New Frontier: Artificial Companions and the Fourth Revolution. *Metaphilosophy* 39(4-5): 651-655.

Goldie, P. (2000). *The Emotions: A Philosophical Exploration*. Oxford: Oxford University Press.

Heerink, M., Kröse, B. J. A., Wielinga, B. J., and Evers, V. (2006). Studying the acceptance of a robotic agent by elderly users. International Journal of Assistive Robotics and Mechatronics 7(3): 33-43.

Hume, D. (2000 [1739/40]). *A Treatise of Human Nature (edited by David Fate Norton and Mary J. Norton)*. Oxford/New York: Oxford University Press.

James, W. (1884). What is an Emotion? *Mind* 9: 188-205.

Levy, D. (2007). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper.

Mori, M. (1970). The uncanny valley (Bukimi no tani. Translated by K. F. MacDorman & T. Minato). *Energy* 7(4): 33-35.

Nagel, T. (1986). *The View From Nowhere*. Oxford/New York: Oxford University Press.

Nussbaum, M. C. (1995). *Poetic Justice: The Literary Imagination and Public Life*. Boston, MA: Beacon Press.

Nussbaum, M. C. (1997). *Cultivating Humanity: A Classical Defense of Reform in Liberal Education*. Cambridge, MA: Harvard University Press.

Nussbaum, M. C. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press.

Prinz, J. (2004). *Gut Reactions: a Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Reeves, B., and Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.

Robins, B., Dautenhahn, K., and Dubowski, J. (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies* 7(3): 509-542.

Shinozawa, K., Reeves, B., Wise, K., Lim, S-H., Maldonado, H., and Naya, F. (2003). Robots as New Media: A Cross-Cultural Examination of Social and Cognitive Responses to Robotic and On-Screen Agent. *Proceedings of the Annual Conference of the International Communication Association*: 998-1002.

Smith, A. (1976[1790]). *The Theory of Moral Sentiments (Edited by A. Strahan and T. Cadell; Reprint edited by D. D. Raphael and A. L. Macfie)*. Oxford: Clarendon Press.

Solomon, R. (1980). *Emotions and Choice. In Explaining Emotions (Edited by A. Rorty)*. Los Angeles: University of California Press, pp. 251-81.

Sparrow, R., and Sparrow, L. (2006). In the Hands of Machines? The Future of Aged Care. *Minds and Machines* 16: 141-161.

Stiehl, W. D., Lieberman, J., Breazeal, C., Basel, L., Lalla, L., and Wolf, M. (2006). The Design of the Huggable: A Therapeutic Robot Companion for Relational, Affective Touch. *Proceedings of AAAI Fall Symposium on Caring Machines*. Available at http://web.media.mit.edu/~wdstiehl/Publications/FS205StiehlWDtoAppear.pdf

Taylor, C. (1992). *The Ethics of Authenticity*. Cambridge, MA: Harvard University Press.

Trilling, L. (1972). *Sincerity and Authenticity*. Cambridge, MA: Harvard University Press.

Veruggio, G. (2006). *EURON roboethics roadmap (release 1.1)*. EURON Roboethics Atelier, Genoa.

DOI: 10.2202/1941-6008.1126

Wada, K., Shibata, T., Sakamoto, K., and Tanie K. (2006). Long-term Interaction between Seal Robots and Elderly People — Robot Assisted Activity at a Health Service Facility for the Aged. *Proceedings of the 3rd International Symposium on Autonomous Minirobots for Research and Edutainment (AMiRE 2005)*. Berlin/Heidelberg: Springer.

Wallach, W., and Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2008.