Are Emotional Robots Deceptive?

Mark Coeckelbergh

Abstract—A common objection to the use and development of "emotional" robots is that they are deceptive. This intuitive response assumes 1) that these robots intend to deceive, 2) that their emotions are not real, and 3) that they pretend to be a kind of entity they are not. We use these criteria to judge if an entity is deceptive in emotional communication (good intention, emotional authenticity, and ontological authenticity). They can also be regarded as "ideal emotional communication" conditions that saliently operate as presuppositions in our communications with other entities. While the good intention presupposition might be a bias or illusion we really need for sustaining the social life, in the future we may want to dispense with the other conditions in order to facilitate cross-entity communication. What we need instead are not "authentic" but *appropriate* emotional responses—appropriate to relevant social contexts. Criteria for this cannot be given a priori but must be learned—by humans and by robots. In the future, we may learn to live with "emotional" robots, especially if our values would change. However, contemporary robot designers who want their robots to receive trust from humans had better take into account current concerns about deception and create robots that do not evoke the three-fold deception response.

Index Terms—Ethics of robotics, emotions, deception, ideal speech conditions, authenticity

1 Introduction

One way to further develop "affective computing" [3], [4] is to build robots with artificial emotions. Justifications provided for designing and employing such robots include not only the argument that it would improve acceptability of all kinds of robots by facilitating human-robot interaction or that it would make them more human-like per se, but also that for some robots, in particular "personal" or "emotional" robots, emotions would be essential to their function. For example, if we wanted to use robots in health care and elderly care, then if what we mean by "care" includes emotional communication, we had better equip these "care robots" with artificial emotions. Moreover, if we wanted autonomous robots to become "moral machines" [8], then given the limitations of "top-down" approaches to morality, it would be wise to develop robots with artificial emotions. Otherwise, I have argued, we might end up with dangerous "psychopathic" robots that only follow rules [1]. For example, if we wanted to use autonomous military robots at all, then unless these robots were "emotional," they would lack true moral capacity and hence present a great moral threat.

Despite these promises, however, affective robotics also invites various ethical concerns. A common objection to the use and development of human-like "emotional" robots is that they are deceptive. For example, it has been argued that the use of robots in elderly care is deceptive. Sparrow and Sparrow warn for cases of *delusion*:

> In most cases, when people feel happy, it will be because they (mistakenly) believe that the robot has properties which it does not (...) It is these delusions that cause people to feel loved or cared for by robots and thus to experience the benefits of being cared for. [5, p. 155]

But are robots with the capacity to engage in emotional communication really "deceptive"? In order to uncover what people mean when they make such a claim, I will analyze this intuitive response, criticize its assumptions, and draw conclusions for robot design.

First, I will distinguish between three claims this vague intuition may refer to: 1) these robots intend to deceive, 2) their emotions are not real, and 3) they pretend to be a kind of entity they are not. Then, I will question if these claims are applicable to robots and show that they rest on problematic assumptions, related to salient conditions for "ideal emotional communication" that are operative in human-robot communication as well as in human-human communication: good intention, emotional authenticity, and ontological authenticity.

After discussing various arguments that may be given in their support, I will suggest that we might want to dispense with the authenticity conditions, but that it is difficult to shed modern-Romantic and Platonic thinking. I will conclude that, in principle and leaving aside other ethical considerations, it is ethically acceptable (if not good) to continue to experiment with "emotional" human-like robots for various reasons, especially since our values might change in the future. However, I will also recommend that today it is wise to build robots that do not evoke the threefold deception response, at least if we wish these robots to receive trust from humans.

Note that there is a broader issue as to whether it is morally acceptable or required to intentionally build robots that can deceive—with or without emotions. For example, recently Wagner and Arkin have developed algorithms which, according to their creators, produce robots capable of deception. They suggest that such robots may be valuable in

• *The author is with the Department of Philosophy, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: m.coeckelbergh@utwente.nl.*

Manuscript received 3 Feb. 2011; revised 22 June 2011; accepted 19 July 2011; published online 18 Aug. 2011.
Recommended for acceptance by A. Beavers.
For information on obtaining reprints of this article, please send E-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFCSI-2011-02-0009.
Digital Object Identifier no. 10.1109/T-AFFC.2011.29.

military operations [7]. As the authors recognize, we need to do more on the ethics of what we may call "deception robotics." However, in this paper I limit my discussion to "deception" as an objection to robots with artificial emotions.

2 "Ideal Emotional Communication" Conditions as Normative Criteria

The intuitive objection that "emotional" robots deceive can be interpreted in various ways. One plausible interpretation I propose here is to analyze the objection into at least three related, but logically independent claims:

1. Emotional robots *intend* to deceive with their "emotions."
2. Robotic emotions are *unreal*.
3. Emotional robots pretend to be *a kind of entity* they are not.

These claims can be regarded as constituting three "ideal emotional communication" conditions: Those who make the deception objection must regard them as conditions we (humans) presuppose when engaging in emotional communication with other entities (including other humans). Let me make explicit a central component of what we may call their "theory of emotional communicative action."

The approach I have in mind here is somewhat analogous to Habermas's "ideal speech conditions" for communicative action. Habermas defined conditions for what he called an "ideal speech situation" in which there is no other force than the better argument (see, for example, [2, pp. 88-99]). According to Habermas, we have to presume the conditions in order to engage in free discourse and reach mutual understanding. Like Habermas's overall philosophical project, this description of the speech situation is centered on rationality. For the purpose of this inquiry, let me create an emotion-centered cousin of his approach, concerned not with speech or rational argument but with emotional communication, in order to make explicit the assumptions of the deception objection. My claim is that the (explicit or implicit) claims of this objection can be understood as involving a theory about conditions presupposed in an "ideal" situation of emotional communication (rather than rational arguments), in which there is mutual understanding based on free emotional interaction (rather than free rational discourse). My basic idea here is that such an "ideal" situation of emotional communication, like rational communication, is highly vulnerable in the sense that it is dependent on the fulfillment of certain conditions. If the conditions are not fulfilled, mutual understanding is not reached and the emotional communicative action is (at least in part) unsuccessful.

Which conditions ? We can reformulate the three claims as follows. When we engage in (cross entity) emotional communication, we must presuppose that

1. The other entity does not *intend* to deceive us.
2. The other entity's emotions are *real*.
3. The other entity does not pretend to be a kind of entity it is not.

When we encounter other entities and engage in emotional communication with them, we must presuppose these conditions. This enables free emotional communication and allows mutual emotional understanding—or at least the *appearance* thereof (see again [1] on the moral significance of appearance.)

Of course, they are "ideal" conditions and hence they may sometimes not be met in practice, but my claim is that proponents of the deception objection must assume that these are three conditions of possibility that enable us to engage in emotional communication with other entities and to reach (the appearance of) mutual understanding.

At the same time, because they are "ideal" conditions, they are *normative* criteria that guide the communication and allow us to evaluate the communication. We can formulate them as being "addressed" to the entity we encounter "before" engaging in emotional communication:

1. Do not intend to deceive us.
2. Make sure your emotions are real.
3. Do not pretend to be a kind of entity you are not.

However, are these conditions and normative criteria applicable to *robots*? And do they hold even for *human* emotional communication? Let me further examine the claims of the deception objection as "ideal" conditions put forward by its related (inexplicit) theory of emotional communication.

3 Good Intention

First, the intention to deceive cannot be ascribed to robots as we know them since intention presupposes consciousness.[1] We must not assume that nonconscious entities can deceive at all. Therefore, those who make the deception objection must reformulate this component of the objection as follows: The robots in question *appear* to have the intention to deceive. They could explain this by saying that the designers or employers of the robot have the intention to deceive users by building or employing their robots in such a way that users communicate with it in a "natural," that is, human-like way. However, then the designer's intention is not so much to deceive but to achieve this kind of human-robot communication. Ideally then, those who communicate with the robot do not feel that the robot "deceives" them. But this is exactly what opponents of emotional robots find objectionable: People (for example, children or elderly people) are *made to believe* that the robot's emotions are real and that it is not a machine. However, this has nothing to do with the *intention* of the *robot*. (I will discuss other criteria below.)

In response, one could change the first condition to "the designer or employer of the robot has no intention to deceive the user." However, it remains questionable if this is indeed a *necessary* precondition for emotional cross-entity communication. Whatever the intention of the designer, we are usually happy to engage in emotional communication

1. Some disagree that intention presupposes consciousness and argue that, for example, animals may count as intentional deceivers without it being the case that they are consciously deceptive. Some robots may also count as "intentional deceivers" under this description. However, if this use of the term intention is intelligible at all, what it means is that we ascribe an intention to a nonconscious entity, a phenomenon which I capture below under the concept of the "appearance" of deception, and in which case the ideal speech condition remains in place.

with another entity as long as it *appears* not to intend to deceive—regardless of the designer's intention. Hence, the first formulation still stands and also holds in the case of human-robot communication: We must presuppose that the robot does not intend to deceive, and the only way to find out this condition is met is to go by appearance.

Moreover, the intention to deceive on the part of the designer or employer of the robot is not *necessarily* morally objectionable: Most people agree that in some cases and situations it is morally permitted to intentionally deceive people in order to improve their well being, as, for instance, in the case of very young children or people with cognitive impairments (elderly or not). For example, no one objects to intentionally showing young children films in which animals like bears or rabbits speak and communicate emotions. Such virtual emotional characters are designed with the intention to deceive and entertain. What matters to the child for engaging with regard to emotional communication with the virtual character is, among other things, that it can safely presuppose that the virtual character—not the designer—does not have the intention to deceive. The same seems to hold for toy robots.

More generally, the assumption that the other entity—biological, artificial, or both—does not intend to deceive me with its (or his or her) emotions seems to be a necessary illusion for sustaining the social life—our relations with humans but also with nonhumans. For instance, it is plausible that people who interact with their biological or artificial pets have to presuppose, as an "ideal emotional communication condition," that the pet does not intend to deceive them with their emotions. If they did not (inexplicitly) presuppose this, communication with their pet would be very hard indeed since they would never feel sure about the meaning of the pet's emotions. The same is true for human-human communication. In general, there is a communication problem when an entity appears to have a bad intention with regard to its emotional communications. However, I suspect that this seldom is the case with "personal" and "social" robots, now and in the foreseeable future, since designers usually want their robots to be accepted by the user. Of course, sometimes there may be a bad intention on the part of the designer or employer of the robot. For example, in elderly care a robot may be used to reduce human contact, in which case there *is* a moral problem (on this point I fully agree with Sparrow and Sparrow). Particular cases deserve further discussion, but the objection against replacing human care workers by robots can hardly be phrased as "the *robot* intends to deceive."

To conclude, it seems that good intention with regard to emotional robots *is* one of the (necessary) conditions for "ideal emotional communicative action," but since current robots cannot have intentions and since designers and employers of robots do not usually have bad intentions[2] with creating or employing "emotional" robots, this part of the deception "objection" does not count as a strong objection—if at all—if it is formulated in these terms. What matters to people is that robots do not *appear* to deceive and then they will be ready to ascribe "good intention" to the robots and engage in communication with them—regardless of whether or not the robots really (can) have any (good or bad) intentions at all.

4 Emotional Authenticity

Second, therefore, the more promising components of the deception objection involve claims about what exactly the deception consists of: that the robotic emotions are not real and that the entity pretends to be an entity it is not. Let me start with the first charge. What does it mean to say that robotic emotions are not "real"? The most plausible answer is that those who object to emotional deception understand "real" to mean "authentic." But what are "authentic" emotions? I suspect that in this case the term refers to at least one of these two normative requirements that need to be met for emotions to be authentic: 1) that the emotions originate in a biological or human entity and/or 2) that the emotions express an inner self or true character.

If the deception objection is interpreted in terms of the first requirement, the charge is that robotic emotions are not real since they are artificially generated. Emotions originate in a biological, organic body rather than an artificial system. This argument assumes that when it comes to emotions, there is a strict natural/artificial or human/nonhuman distinction and that this distinction has normative authority when it comes to determining the reality or authenticity of emotional communication. This is a common argument. For example, we accept that some animals have "real" emotions, but we deny this to "machines" *because* they are machines, because they are artificial.

For the sake of argument, let us accept the first assumption and suppose that there is a strict distinction between the natural (the biological) and the artificial. That leaves us with the latter assumption, which concerns the authority of the distinction. This assumption should at least be questioned. Why would the natural-artificial distinction have the strong normative authority it is supposed to have by those who employ the deception objection, given that most of us accept emotional engagement with *artificially* generated characters in films and games that merely *appear* biological? What is its authority if some of us accept emotional engagement with *robots*—virtual or real—that appear to have emotions? Is it *morally objectionable* if these people do this?

It seems to me that in the latter case, if the human engages in "emotional" communication with the robot, it would be *correct* to call the robot's emotional communication "deceptive" if we mean by this term "artificially generated," but it would not be appropriate or relevant to do so in that context. While it is *true* that the robot's emotions are artificially generated, to employ the real/unreal or authentic/nonauthentic distinction to call them "deceptive" would not be an appropriate response in this situation and context. Does it *matter* that the emotions are artificially generated if people enjoy communication with the robot?

Thus, *if* its underlying assumptions are right, the emotional authenticity objection holds. Perhaps robotic emotions are not "real," if that means that they are artificially generated. But even if the assumptions were right and

2. Note that this absence of bad intentions does not dissolve them from responsibility for the (side) effects of their actions.

support this part of the deception objection (the emotions originate in an artificial entity), in these situations and contexts that aspect of the objection does not matter. (Of course, if the emotional communication hampers, then the user or an observer may employ this objection in order to call attention to and/or explain the hampering— although it would need to be established that the artificial origin of the robot's emotions, rather than another factor, really is the cause of the hampered emotional communication.)

Another interpretation of the emotional authenticity condition would be to require that the robot's emotions express its "inner self" or its "true" character. This interpretation is based on an analogy with human-human communication: We tend to assume that someone's emotions somehow reflect his or her inner self or true character. Therefore, this interpretation of the emotional authenticity condition raises at least two problems.

First, as most opponents of emotional robots would agree, robots do not have an "inner self" or "true character" as opposed to the "self" or "character" it displays in communication. (Of course, it could learn or be programmed to have multiple characters, but given their identical origin, none of these characters would have the status of being more "inner" or "true" than others.) Someone who nevertheless wants to hold on to this interpretation therefore must presuppose that—whatever its precise origin (preprogrammed, learned, emergent, or a combination of these)—the robot has a *virtual* "inner self" or *virtual* character as the virtual soil in which the robot's artificial emotions are rooted. In that case, the authenticity of the robot's emotions could be assessed provided one had access to that virtual self and artificial character. However, I guess this is not what opponents of emotional robots mean when they call robotic emotions deceptive in the sense of inauthentic. If they endorse the "inner self" interpretation of emotional authenticity at all, they probably rather mean that the robot has no inner self or character at all, whereas humans express emotions that *are* rooted in such an inner self or character, and that therefore robots are "inauthentic" *in principle*. But is this assumption about humans justified?

Second, then, we must also question the assumption that *human* emotions are expressions of an inner self or true character. In order to look criticaly at this common intuition, let me (re)construct the opposite claim that there is no such self or character. First, we must consider the claim that there is *no self*. Eastern traditions of thinking are known for lacking the assumption of a self that lies at the bottom of experience and thought, but the assumption has also been questioned by Western science. For example, in the field of cognitive science, Varela has referred to Buddhism, Taoism, Confucianism, and indeed to contemporary science to argue that at most we have a virtual self [6]. Second, while the ancients already used the term "character" as a collection of traits (although the Greeks first used it to mean the impression of a mark on a coin), in the Western tradition the idea of an *inner* self (and the related idea of a fixed personal character) is relatively recent—the ancients lacked this idea. Moreover, the normative claim that we should try to achieve emotional authenticity is a modern, Romantic belief that was first held only by a limited number of people (e.g., artists, writers, philosophers) and has arguably only come to full bloom today. In other words, the belief that humans have an inner self is more *contingent* and less universal than assumed in this interpretation of the deception objection: True or not, it belongs to a particular, cultural-historical way of thinking and cannot simply be taken for granted. Therefore, it cannot be used to characterize and insist on the authenticity of human emotions as against the supposed inauthenticity of robotic emotions without further discussion.

Again, those who still wish to cling to the inauthenticity objection may adopt a different definition of authenticity and say that it requires us to "be true to your virtual self" rather than to "be true to your real, inner self." This is an interesting interpretation of authenticity which deserves further discussion. But, as said, most defenders of the inauthenticity objection rather seem to start from the premise that robots have no self or character at all when they argue that robotic emotions are inauthentic.

5 Ontological Authenticity

A third part of the deception objection concerns the charge that "emotional robots" pretend to be an entity they are not. In particular, the accusation is that they pretend to be human or that they pretend to be a (nonhuman) animal with emotions, and that this kind of pretending is morally unacceptable. The related "ideal emotional communication" condition is then that when we engage in emotional communication with other entities, we must presuppose that the other entity does not pretend to be an entity it is not. Like the other conditions, this "ideal" condition is understood to function as a normative requirement that is supposed to guide our dealings with other entities: The other entity *should* not only have good intentions with its emotions and have authentic emotions, it *should* also have what we may call *ontological* authenticity.

The argument for seeing ontological inauthenticity as morally problematic rests on the view that seeing the world in the "right" way is morally required. For example, Sparrow and Sparrow claim that "failure to apprehend the world accurately is itself a (minor) moral failure. We have a duty to see the world as it is" [5, p. 155]. Is there such a duty? Whether or not there is, to require this from people assumes that we *can* make a sharp distinction between "the world as it is" (reality) and the world as it appears (illusion). But it is not obvious that this assumption can be maintained. The Platonic metaphysics presupposed here implies that there is a sharp distinction between what the entity really is and how the entity appears to us. The part of the deception objection under consideration here argues 1) that in emotional communication a robot may *appear* to us as being the kind of entity that can have emotions, e.g., as human, but that *in reality* it is a mere machine, and 2) that this kind of "pretending" is morally unacceptable.

According to an alternative philosophical-epistemological tradition (phenomenology), however, making such a sharp distinction between reality and appearance is impossible: Our view of what is real is *always* mediated or constructed, what we consider to be real is reality-as-it-appears-to us. In Heideggerian language, we know the world as *beings-in-the-world*; we are not detached Cartesian

egos that can observe reality from a view of nowhere. If this is right, then to invoke a reality-illusion distinction in order to criticize emotional robots is to construct these robots as machines as against another appearance: in one context the robot appears as a machine (e.g., in a scientific context or in the context of detached philosophical reflection), but in another context the robot appears to the person as "more than a machine." But from this perspective, it is not obvious why one of these two appearances would have ontological priority. It seems that the robot can appear to different people in different ways at different times and in different contexts (e.g., the context of home care and the context of a scientific lab). The robot has different Gestalts, which cannot be experienced at the same time, but which are both "real" possibilities. Thus, the charge of ontological inauthenticity requires a particular kind of "work": It requires that the opponent of emotional robots lets these robots *appear* as "mere machines" or *constructs* these robots as "mere machines" and thereby dismisses and excludes other appearances—for example the appearance of the robot-as-companion. If this is true, then "deception" cannot be used against emotional robots, if by deception we mean "ontological inauthenticity," and we should not include it in the list of "ideal emotional communication conditions."

Furthermore, even if we maintain standard Platonic epistemology, it remains questionable how *relevant* it is to invoke the real-illusion distinction, understood as an ontological distinction, *in the particular context under consideration*. Is it always relevant to tell someone who loves his robot that the robot is only a machine? Perhaps it is relevant in some particular cases, for example, when these experiences lead the robot lover to reject all human contact or all human friendship. But such (sad) cases should not be generalized and are likely to be the exception rather than the rule.

Moreover, due to the very nature of their inexplicit epistemological project, opponents of emotional robots find themselves in the position where they have to tell people who use these robots that they are being deceived: Whereas *they* (lay people) actually experience well-going emotional communication with the robot, they should not since the robot really is deceptive since the robot really is a mere machine (according to expert-philosophers). This renders the opponents at least as "paternalistic"[3] as their (largely imaginary) opponents, who see no problem at all in this "deception" and could be accused of "paternalism" as well since they justify giving robots to people (e.g., in elderly care) for their own good, regardless of what these people want. In other words, in the end the deception charge risks not only irrelevance but might also itself attract moral criticism insofar as it is a top-down, "expert" view condemnation of the emotional engagements of lay people who are told by the expert-philosopher that their experience is illusionary. Such a manner of doing may not be morally wrong per se, but like all paternalisms it requires further justification.

This is not to say that there are no ethical problems at all with emotional robots, but the discussion presented here suggests that to call them "deceptive" is philosophically problematic in several ways.

6 Conclusion

Where does this inquiry lead us? Let me further reflect on the three parts of the deception charge and their assumptions: good intentions, emotional authenticity, and ontological authenticity. I have identified the three parts as conditions for an "ideal emotional communication" situation or normative criteria for cross-entity communication, assumed by those who object to emotional robots. I have shown that they in turn incur various philosophical assumptions which have been questioned and discussed. What can be concluded for now?

The assumption that when engaging in emotional communication the other entity has good intentions with this communication might be a bias or illusion we really need for sustaining the social life. It seems to be an indispensible "ideal" condition that must be presupposed for *any* communication in order for it to contribute to mutual understanding (or the appearance thereof): between humans and robots but also between humans. But what about the other conditions? Allow me to explore the following thought (which is more a speculation than a conclusion from the previous discussion): Given the problematic character of the other, "authenticity" assumptions, in the future we might want to dispense with these conditions in order to facilitate cross-entity communication. But what would it mean if we were to shed modern-Romantic and Platonic thinking on this matter? What could possibly replace authenticity as a normative criterion for emotional communication?

Let me suggest that what we need instead, if anything, are not "authentic" but *appropriate* emotional responses—appropriate to relevant social contexts. Criteria for this "appropriateness" cannot be given a priori but must be learned and adapted if necessary—by humans and by robots—and they must be defined for, and enacted in, the situation and context in question. This view presupposes a thoroughly relational view of humans and other entities, which holds that it makes no sense to consider entities in isolation from their natural-social environment. Whether or not we support this view by Heideggerian phenomenology and theories of embodied cognition, it also assumes that emotional communication is not an exchange between two entirely separate entities with an "inner" core from which individual emotions spring, but is something that is already a *communal* and *public* process from the start and that constitutes what intelligent entities like us consider to be their "inner life," "true character," and perhaps also their very self.

This view's contextualism does not prevent us from making generalizations, at least if they keep in touch with practice and as long as they are not taken to be perennial truths that have no relation to changing technological and moral worlds. For example, I guess that in the present

3. When I use the term "paternalism" here I mainly refer to a rather "weak" form of paternalism: the idea that a person A (expert, philosopher) knows better than person B (lay person) what is good for B because A knows the truth and is therefore justified to *tell* B what to do and how to live. I do not mean a stronger form of paternalism that concludes from this that A (or a state informed by A's view) is justified to *force* B to act and live in a certain way against B's will.

context of Platonic-Romantic modernity, it is best that nonhuman artificial entities capable of, and designed for, emotional communication, communicate their emotions in a way that appears to humans as expressions of what they "really feel inside" and what they "really are." (Note that if we call these communications "expressions", we are already assuming that the entity brings out, reveals something that is inside it.) Hence, in these times it is understandable that robots which appear "inauthentic" or "superficial" in the sense of showing emotions without connection to what they really feel or are will be seen as "deceptive." However, such generalizations have their limitations. Even Romantic modernity does not require that all emotional communication be "deep" at all times in all situations and contexts— with deep meaning emotionally authentic as defined above. For example, in many contexts a "friendly" appearance (e.g., friendly facial expressions like smiling, a body that does not look threatening, etc.) is enough. (Consider, for instance, intelligent service robots that would need to provide only minimal emotional communication in order to function smoothly in a human, social environment.) How "deep" the emotions need to appear would depend on how "personal" the robot's function is and the particular form its emotions would need to take on how the robot has learned to behave in the particular social context.

To conclude, experimenting with "emotional" robots helps us to reflect on ourselves as humans. For example, it makes us reflect on what we mean by emotional deception and authenticity, on the nature of emotional communication and its "ideal" conditions, and on what kind of emotional communication is needed in which context. The discussion in this paper urges defenders of the deception objection to express more clearly what exactly they find objectionable. Ultimately, we have to further discuss what it is that we value in humans and in human contact.

On a more practical note, one thing that may be problematic about emotional robots is that there is a gap between human expectations and what current robots can deliver when it comes to social and emotional communication and human value. But, to address this problem, it is not very helpful to accuse robots or their designers of deception. Rather, a more fruitful approach is fine-tuning human expectations about robots; it seems that we expect too much of them. Companies should be careful when making claims about the robots they produce and robot designers can also contribute to this task. As Picard wrote: "a machine may need to explain what it can and cannot do" [4, p. 114].

Moreover, whereas now we tend to view emotional communication with (personal) robots from a Platonic and Romantic perspective, in the future we may well learn to live with robots we now call "deceptive" if our values change and if we feel more secure in our relations with other entities. Relating to robots is a learning process.[4] I do not know if such a value change would be good; this is difficult to judge since we only have our current values to go by. We can only try to *imagine* possible future worlds—technological-moral worlds. However, with regard to contemporary robotics, it is recommended that designers of "personal" or "social" robots who wish their robots to receive trust from humans had better take into account current concerns about deception and build robots that do not evoke the threefold deception response. It may take a while before we move beyond Plato and Rousseau, if we ever do.

4. There may be limits to the speed of these changes due to the evolutionary origins of our emotions and other aspects of our psychological make-up. For example, in the short term, it may be difficult to reshape our "sense" of deception—if this is desirable at all. On the other hand, learning may have a deeper impact on us than we might think and we should not forget that the same evolution also made us into the social animals we are, which may facilitate rather than hinder this learning process and help us to relate better to robots.

References

[1] M. Coeckelbergh, "Moral Appearances: Emotions, Robots, and Human Morality," *Ethics and Information Technology*, vol. 12, no. 3, pp. 235-241, http://www.springerlink.com/content/103461/, 2010.

[2] H. Jürgen, *Moral Consciousness and Communicative Action*, T.C. Lenhardt and S.W. Nicholsen, eds. MIT Press, 1983.

[3] P. Rosalind, "Affective Computing," MIT Technical Report No. 321, 1995.

[4] P. Rosalind, *Affective Computing*. MIT Press, 1997.

[5] R. Sparrow and L. Sparrow, "In the Hands of Machines? The Future of Aged Care," *Minds & Machines*, vol. 16, pp. 141-161, 2006.

[6] F.J. Varela, *Ethical Know-How: Action, Wisdom, and Cognition*. Stanford Univ. Press, 1999.

[7] A.R. Wagner and R.C. Arkin, "Acting Deceptively: Providing Robots with the Capacity for Deception," *Int'l J. Social Robotics*, vol. 3, no. 1, pp. 5-26, 2011.

[8] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford Univ. Press, 2008.

Mark Coeckelbergh is an assistant professor of philosophy at the University of Twente, The Netherlands. His publications include *Liberation and Passion* (2002), *The Metaphysics of Autonomy* (2004), *Imagination and Principles* (2007), and numerous articles on ethics and technology. His current research focuses on the philosophy of information technology, robotics, and artificial intelligence.