Ethics Inf Technol (2010) 12:235–241
DOI 10.1007/s10676-010-9221-y

Moral appearances: emotions, robots, and human morality

Mark Coeckelbergh

Published online: 17 March 2010

Abstract Can we build 'moral robots'? If morality depends on emotions, the answer seems negative. Current robots do not meet standard necessary conditions for having emotions: they lack consciousness, mental states, and feelings. Moreover, it is not even clear how we might ever establish whether robots satisfy these conditions. Thus, at most, robots could be programmed to follow rules, but it would seem that such 'psychopathic' robots would be dangerous since they would lack full moral agency. However, I will argue that in the future we might nevertheless be able to build quasi-moral robots that can learn to create the appearance of emotions and the appearance of being fully moral. I will also argue that this way of drawing robots into our social-moral world is less problematic than it might first seem, since human morality also relies on such appearances.

Keywords Robot morality · Human morality · Emotions · Rule-following · Mental states · Feelings · Appearance

Introduction: morality and emotions

Can robots be moral? One of the first attempts to think about 'robot morality' was presented by Asimov in his robot stories. His 'Laws of Robotics', introduced in the story *Runaround*, prescribe the following rules to robots:

1. First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. Second Law: A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
3. Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (Asimov 1942).

These rules illustrate a common way of thinking about robot morality which hinges on several problematic assumptions.[1] In this paper, I focus on the assumption that morality is about following and applying rules and has nothing to do, or should have nothing to do, with emotions. This view does not come as a surprise: many moral theories—usually applied to humans—suggest this or are interpreted in this way. For instance, consequentialist ethics fashions abstract rules, such as the 'no harm principle', that are supposed to cover our moral intuitions and the difficult moral situations in which we might find ourselves.[2] And Kant's ethics is often interpreted as supporting a view of morality as the application of rules.[3] However, in the traditions of virtue ethics, Humean ethics, and pragmatist ethics we can find alternative conceptions of morality that put less emphasis on rule following and

M. Coeckelbergh (✉)
Department of Philosophy, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: m.coeckelbergh@utwente.nl

[1] For instance, the Laws seem to limit the range of possible human-robot relations to the master–slave model.

[2] For contemporary examples of such rules and arguments see Peter Singer's work.

[3] I do not agree with this interpretation of Kant. The categorical imperative is not a rule but at best meta-rule asking from us to reason from the moral point of view when we make rules (when we, as autonomous beings, give the rule to ourselves). But as I argued in my book […] this leaves open a lot of space for types of moral reasoning that require the exercise of imaginative and emotional capacities.

rule application and more on the role of emotions and imagination. What we may call the strong 'emotion' view of morality makes a twofold claim: a descriptive one and a normative one. First, it is held that emotions do, as a matter of fact, play a role in human morality. This 'weak' claim is descriptive and ethicists from the rule-following tradition usually agree with it. *That* emotions play a role in morality has been demonstrated through psychological and neurological research. For instance, Greene and others have conducted fMRI investigations of emotional engagement in moral judgment (Greene 2001) and previously Damasio has cited neuroscientific studies to defend the importance of emotions in morality (Damasio 1994). And of course Kant and the utilitarians already observed and acknowledged that feelings play a role in morality. However, Kantians and many other modern philosophers tend to disagree with the further, strong *normative* claim that the role that emotions play should be evaluated positively. For instance, the work of Martha Nussbaum, which is highly representative of this stronger view, draws together the various alternative traditions mentioned above. Inspired by Aristotle, Stoicism, Adam Smith's theory of moral sentiment, and Jamesian pragmatism, Nussbaum has articulated a vision of morality that depicts emotions as indispensable for moral judgment and constitutive of the good life and human flourishing (Nussbaum 1990, 1994, 1995, 2001).[4] For example, based on a neo-Stoic descriptive account of emotions, in *Upheavals of Thought* she argues that emotions are not blind forces but intelligent responses that teach us what is of value and importance and that therefore they are not marginal but central to adequate ethical reasoning (Nussbaum 2001).[5] Moreover, many of us intuitively feel that someone who only follows the rules without emotion is not only lacking in moral capacity but is also insane or even dangerous. Consider discussions of psychopathy, it is suggested that psychopaths can follow rules but do not have the capacity to *feel* that something is morally wrong. Even though they may act in accordance with moral conventions, they lack the appropriate emotion that motivates non-psychopaths to act right. As Kennett remarks, one can imagine a psychopath who knows the conventions but 'still fails to be moved by moral concerns' (Kennett 2002, p. 342). If this is 'moral reasoning', it is only so in a very awkward sense. Although it is certainly reasoning, on the 'emotion' view it could not be called 'moral': at most, it is conventional reasoning.[6]

From the standpoint of this strong 'emotion' view, it would not only be wrong to call such rule-following robots 'moral', but it might also be dangerous to build them—after all, they would be 'psychopathic' robots. They would follow rules but act without fear, compassion, care, and love. This lack of emotion would render them non-moral agents—i.e. agents that follow rules without being moved by moral concerns—and they would even lack the capacity to discern what is of value. They would be morally blind. If these robots were given full independence—absence of external control by humans, which is another condition for full moral agency—they would pose danger to humans and other entities.[7]

A full defence of these positions would require a longer work. But let us assume for the sake of argument that (1) the 'emotion' view is right about the positive relation between emotions and morality and that (2) we wish to avoid 'psychopathic' robots as described above: intelligent autonomous robots that can follow rules but lack emotions and are therefore—on the 'emotion view—amoral robots. Thus, if we want to build *moral* robots, they will have to be robots *with emotions*. But can such robots be built? In what follows, I first argue that it is not likely that in the foreseeable future we will be able to build such robots, because to do so these robots would have to be *conscious*, they would have to have *mental states*, and we would have to be able to *prove* that they have these things. However, I then argue that we might nevertheless be able to build quasi-moral robots that produce the appearance of being moral, and that this may suffice for our purposes since human morality also depends on such mere appearance.

Robot emotions?

Let me first note that *if* it were possible for robots to have emotions, then emotions need not be 'given' to robots as a kind of ready-made cognitive package but could be developed by the robots themselves. In *Moral Machines* Wallach and Allen have voiced their criticism of Asimov and related attempts to build 'moral machines' as rule-

[4] Note that there are tensions between the theoretical traditions mentioned here, for instance between a Human and a virtue ethics approach (see for instance Foot's criticism of Hume, Foot 2002), but Nussbaum has managed to reconcile them in an attractive way.

[5] Influenced by the Stoics, Nussbaum writes that emotions are not just 'unthinking forces that have no connection with our thoughts, evaluations, or plans' like 'the invading currents of some ocean' (Nussbaum 2001, p. 26–27) but, by contrast, more like 'forms of judgment' that 'ascribe to certain things and persons outside a person's own control great importance for the person's own flourishing.' This renders emotions acknowledgments of vulnerability and lack of self-sufficiency (Nussbaum 2001, p. 22). Note also that this view is not Stoic but neo-Stoic since Nussbaum rejects their normative view of the role emotions should have (the Stoics evaluated the role of emotions negatively) and revises their account of cognition.

[6] Note that emotional moral reasoning does not exclude taking into account rules, laws and conventions.

[7] Given the role of emotions in making moral discriminations, we would not even want 'psychopathic' military robots.

following machines (Wallach and Allen 2008, pp. 83–98). After discussing various limitations of what they call a 'top–down', rule-following approach to ethics,[8] the authors discuss 'bottom-up' and developmental approaches to morality (pp. 99–116). The idea is that to become a moral, 'decent' human being one needs moral development, which depends on complex nature-nurture interactions (pp. 99–116), and their idea is: why not apply this to robots? Robots could mimic child development and be 'raised'.[9] If robots were to have the capacity to learn and evolve as well, then there would be no need to 'give' emotional capacities to robots: on this view, 'all' that robot designers need to do is to equip their creations with the capacity to learn and/or evolve into emotional entities. Then we could hope that they will develop into moral machines, as we humans did and do (as a species and as individuals).

However, is it possible for robots to have emotions at all—regardless of their origin—or can they only *imitate* emotions? In order to explore this question we need to engage with emotion theory. The literature on this topic is vast, but let me limit my discussion to two influential views, which I shall call the cognitivist theory and the feeling theory. According to cognitivists, emotions are propositional attitudes, beliefs, or even judgments (for example de Sousa 1987; Solomon 1980; Nussbaum 2001; Goldie 2000). Nussbaum's neo-Stoic view is a good example: she tends to agree with the Stoics that emotions are judgments (Nussbaum 2001, p. 22). Feeling theories of emotions, by contrast, understand emotions as awareness of bodily changes (for example James 1884; Prinz 2004). William James argued that the feeling of bodily changes *is* the emotion; the mental state follows those changes instead of preceding them (James 1884, p. 189–190). To put it crudely: for cognitivists emotions are more a matter of 'mind' than of 'body', whereas for feeling theorists emotions are more a matter of 'body' than 'mind'.

According to either of these emotion theories, can robots have emotions? There are at least two problems here. Both theories assume that one of the necessary conditions for any entity to have emotions is that that entity has mental states and/or consciousness—i.e. that it (1) has the *capacity* for having mental states (or consciousness) and (2) actually and *really* has these mental states when having emotions (and is actually and really conscious at the time). Although they might not use the same definition of 'mental states' or 'consciousness', cognitivist theories and feeling theories make these assumptions too. For cognitivist theory, if we did not have mental states and consciousness, we could never have emotions-as-attitudes, emotions-as-beliefs, or emotions-as-judgments. (This is because either having mental states and conscious are *conditions* for attitudes and beliefs or because those attitudes and beliefs *are* mental states.) For feeling theory, emotions depend on the capacity to have mental states as well since without such states we could not become aware of our bodily changes. Thus for both theories, it turns out, mental states and consciousness are necessary conditions for having emotions or emotions are themselves mental states.

If this is true, we must conclude that *if* robots do not have mental states and do not have consciousness, robots cannot have emotions: they are unable to form attitudes or judgments and they are unaware of their 'bodily' changes. Not only are they 'mindless' since they are unconscious; in a phenomenological sense they are also 'bodiless'. While we might agree that they have a 'robot body' (that is, they have one in our eyes), the robots are unable to perceive *themselves* as 'being' their body or as 'having' their body, to use Merleau-Ponty's terms (Merleau-Ponty 1945). Thus, this argument would exclude all robots that are being built at present and that will be built in the foreseeable future from the category 'emotional robots'.

But what if in the future someone built a conscious robot? Here we encounter a second problem. Suppose that someone *claims* to have built a conscious robot that has mental states or consciousness. Suppose even that a *robot* makes such a claim about itself: 'I am conscious'. Then how can we find out if this is true? One could use some kind of Turing test,[10] of course, but this does not give us *certainty* that the entity is conscious. After all, we are not even absolutely certain that other humans are really conscious. I will further develop this sceptical point below and

[8] The authors argue that trying to build robots according to the rule-based model (that is, turning the rules into algorithms and build them into robots) cannot succeed since such 'commandment' models face the problem of conflicting rules. Overriding principles based on moral intuitions we have do not solve this problem since they might not even be universally shared within one culture (Wallach and Allen 2008, p. 84). Moreover, applying deontological and consequentialist theories requires one to gather an enormous amount of information in order to describe the situation and in order to predict, which may be hard for computers—and indeed for humans (p. 86). They give further reasons why morality is hard to computate, which is particularly problematic for Bentham-type utilitarian approaches to ethics (pp. 86–91). They also explicitly discuss problems with Asimov's laws (pp. 91–95) and, more generally, problems with deontological abstract rules, which run into similar problems as consequentialist theories since this approach also requires us to predict consequences (pp. 95–97). These problems do not only get roboticists into trouble; they cast doubt on the ambitions of much normative moral theory: it shows that (top–down) theory is valuable but that it has significant limitations. .

[9] Today there are already robots that have some capacity to learn in and from social interaction, for instance the robot Kismet developed by Cynthia Breazeal at MIT. In a sense, she has 'raised' the robot. However, these developments do not approach human moral and emotional learning.

[10] The Turing test has been proposed by Alan Turing to test if an entity is human or not (Turing 1950).

respond to it by biting the bullet: I will argue that our social and moral life depends on appearance.

However, before I continue, let me pause to consider a behaviourist objection. Behaviourism does not assume that having mental states is a necessary condition for having emotions. On the behaviourist view, emotions are to be regarded as behaviour rather than something that goes on 'in the mind'. If that is the nature of emotions, surely then emotional robots (and thus *moral* robots) are possible (or, more precisely, at least one necessary condition for moral agency would be fulfilled). This view is perhaps less intuitively attractive, because it opposes the common sense view that emotions are something 'inside', but that does not necessarily mean that this view is untrue. So, is it true?

I propose a different response to the problem, which also answers the behaviourist objection. In the next section, I explore a view that also shifts the focus to the 'outside' rather than the 'inside', but is not behaviourist but instead phenomenological, where the emphasis is not on the behaviour of the robot but on what that robot does to us, in particular, how it appears in our (human) consciousness. In this way, three goals are achieved. First, it avoids human consciousness and human subjectivity being disregarded and being replaced by behaviour. Second, because of this, the account can do justice to the phenomenon that humans talk about robots as if they had emotions. Thus it saves both the claim that humans have mental states and the claim that robots appear to have emotions. Third, it provides an answer to behaviourism. Behaviourism—at least in its ontological version, not in its methodological version—denies both claims since it removes the inner life from its vision. But the ability to have some kind of inner life (whatever its ontological status) is a condition for perception, for appearance, and for the 'observations' of the behaviourists. Ontological behaviourism, therefore, is self-refuting, and at best, it is a method to study appearances.

The importance of appearance

The capacity to have mental states and consciousness are not only conditions for having emotions; they are also generally regarded as conditions for moral agency and moral responsibility. Arguments for considering robots as moral agents rely on the robot having mental states and being conscious. For instance, theories of human responsibility require that agents have control over their actions and that they know what they are doing.[11] Fulfilling these conditions requires the ability of having mental states and of being conscious: the control condition is usually interpreted in volitional terms—that is, some kind of inner state—and 'knowing what one is doing' requires that one is conscious. Moreover, further to the assumptions made at the beginning of this paper, having emotions is itself a criterion for moral agency. If we require that moral robots have emotions, then we meet the same problem twice. In order to show that a particular robot has emotions and therefore fulfils one of the criteria for *moral agency*, we have to provide proof that a robot is capable of having mental states and consciousness. And the same proof is required since these conditions for having *emotions* are themselves direct criteria for moral agency. But as I previously asked, how can we really know if a robot—or for that matter a human—has such a mental state? The robot may fake the mental state. As long as we hang onto the 'reality demand', we might try to build robots with mental states, but we can never know for sure if we succeeded. Thus, given that current robots lack the capacity to possess genuine mental states and consciousness, and given the more general scepticism about how we could ever establish that *any* being has these capacities, current robots cannot be considered as having emotions and the prospects for designing such robots in the foreseeable future are dim.

However, why put such high demands on robots if we do not demand this from humans? Our *theories* of emotion and moral agency might assume that emotions require mental states, but in social-emotional *practice* we rely on how other humans appear to us. Similarly, for our emotional interaction with robots, it might also be sufficient to rely upon how robots appear to us. (Note that this is not a matter of (robot) behaviour alone. As indicated above, even behaviourists need to possess their own consciousness, so that the robot's 'behaviour' can have the appearance to them of exhibiting consciousness.) As a rule, we do not demand proof that the other person has mental states or that they are conscious; instead, *we interpret* the other's appearance and behaviour as an emotion. Moreover, we further interact with them *as if* they were doing the same with us. The other party to the interaction has virtual subjectivity or quasi-subjectivity: we tend to interact with them as if our appearance and behaviour appeared in their consciousness.[12] Thus, if robots were sufficiently advanced—that is, if they managed to imitate subjectivity and consciousness in a sufficiently convincing way—they too could become the quasi-others that matter to us in virtue of their appearances. As emotional and social beings, we would come to care about how we would appear to

[11] These conditions have already been proposed by Aristotle and are endorsed by many contemporary writers on freedom and responsibility.

[12] More generally, there is a kind of virtual intentionality (understood in a phenomenological sense): it appears as if the other is conscious and as if that consciousness is directed to objects.

robots—about what robots would 'feel' and 'think' about us. Thus robots would become virtual subjects or quasi-subjects with virtual emotions or quasi-emotions.

This phenomenological description and interpretation is not only plausible as a way of making sense of observations of human-robot interaction (such studies are important and we can learn a lot from them); we can also observe and experience it in human–human interaction. If robots resembled humans in many ways, we could expect that they would be regarded and treated in the same way as we treat (other) humans and that we would adapt our own actions and thinking accordingly, as we do when interacting with other humans. Of course, to the extent that existing and future robots do not live up to their designer's ambitions to create convincingly human-like robots, they may be regarded as 'mere things' rather than quasi-subjects. But the point I make here is that this depends on appearance of the robot to humans, not on whether the robot actually has mental states or consciousness.

If this is true, then what might we conclude about (the design of) robot emotions? Is it any 'easier' to design robots that *imitate* emotions (and therefore moral agency) than robots which really have emotions? Perhaps it is. But whether or not it is easier, for designers the advantage of taking this perspective on robot emotions is that they do not have to worry about creating 'internal states' or consciousness; instead they can continue their job as many of them see it: as work of imitation rather than creation.[13] If they continue along this path, they might create robots that learn to produce the appearance of being fully moral, including the appearance of emotions-as-cognition and emotions-as-feelings. Such robots would appear to have beliefs and the ability to judge. They would also appear to respond with feeling to what they perceive in the 'external world' and in their robot 'body'. They would, indeed, appear human.

But whatever we conclude for robots, the other important conclusion of this discussion concerns *human* morality: to the extent that human morality depends on emotions—both in its conditions (having the capacity) and in exercising these capacities—it does not require mental states but only the *appearance* of such.

Disability, slavery, and property

The argument about emotions and moral agency can be expanded to moral status. Before concluding, I will briefly discuss this topic in order to further complement the picture of morality that emerges here. So far I have mainly discussed robots as moral *agents*, that is, robot morality was about how robots should act. But the scope of our moral world is larger since it also concerns entities as moral patients: objects of moral concern. Can robots be moral patients and under what conditions? I partly answered this question in the previous section: if they appear to us as humans, then we will treat them as such without requiring them to actually possess (or *prove* that they possess) mental states, which is now generally the case when we define the scope of human morality. Fortunately appearance has usually been considered sufficient to draw entities into our moral world, and we do not require proof of mental properties to grant them moral status. We have other conceptual tools available for this purpose and often we are content with appearance. Let me give three examples of issues in the domain of moral status: disability, slavery, and property.

Disability

Not all humans have (the capacity for having) mental states, not all humans can have the mental states required for emotions and other morally important capacities, and not all humans who *do* have these mental states are able to develop and exercise their moral-emotional capacities. Nevertheless, we include these humans in our moral world. Someone who suffers from a cognitive disability, for instance, is not expelled from our world of moral concern. Many of us would not even expel humans who do not have mental states at all and others at least consider it a moral question whether or not to grant moral consideration to such beings. Thus, while the capacity for having emotions might be necessary for moral agency, that capacity is not necessary for moral patiency. Instead, we try to find out whether or not the entity in question is human. The application of this criterion relies on appearance: when such a human being *looks like* a human being, we give it the benefit of our moral concern (which might be articulated in terms of human rights or other moral concepts). Thus, we rely on a 'speciecist' foundation (to use Singer's term): it is sufficient that one is—that is, *appears*—human. Now if this is true, that is, if our justification of moral concern for humans relies on appearance, it would be unreasonable to demand that robots have mental states in order for them to be objects of our moral concern. Instead, it is more likely that as robots become more advanced (i.e. more autonomous, intelligent, etc.), we will develop separate moral categories for different robots, as we do for animals: we treat some animals differently, which is not based on a mental states condition but on appearance. For example, we treat a particular dog as a pet since it appears

[13] Perhaps this helps to interpret the phenomenon that Japanese designers are more advanced at making humanoid robots: they tend to understand themselves as imitators of nature rather than creators ('playing God'), which appears to be more a Western idea.

to have those emotions that make us see it as a companion. Whether or not this way of proceeding in our moral reasoning is morally acceptable, it is the way we usually 'do' morality. It would be inconsistent to demand proof that robots have mental states or consciousness if we do not demand this of some severely disabled humans or animals that may even have capacities that exceed those of some disabled humans.[14]

Slavery

What moral categories does our Western tradition provide for robots? A category which combines some degree of moral agency with some degree of moral patiency is that of the slave. Let me try to reconstruct historical reasoning about slaves. Slaves were not defined as human, yet they had what must have been considered as the *appearance* of humans or human-likeness. In many ways they were comparable to the category of intelligent 'work animals': we considered such animals 'thingly' enough to use them for our purposes and not grant them their own lives, yet we found them 'human-like' enough to grant them a minimal degree of moral agency and moral patiency. We granted some moral agency and patiency to such animals since in our perception they had what we might call 'virtual' emotions or quasi-emotions. Although we would not have acknowledged that they really had emotions and the mental states required for having emotions, we might have been content to act towards them *as if* they have emotions. We did this on the basis of appearance. Similarly, some robots might be regarded in the same way as those work animals that have the appearance of emotions. In a similar argument for treating robots well, the apparent presence of feelings gives rise to moral obligations on the part of the humans for the treatment of the robot and to moral expectations towards the robot. While I do not wish to argue for or against this way of regarding humans, animals or robots, I wish to draw attention to moral vocabularies we can draw on to frame 'moral' robots and emphasize that this is a language which does not require proof of mental states. This, at least, is how (Western-style) human morality has worked, how it has shaped the lives of humans and non-humans.[15]

Property

If we care about giving robots some moral consideration but hesitate to grant them the moral status of slaves or working animals, there is also an indirect argument for moral consideration. If we regard robots as things, we have some obligations to treat them well in so far as they are the property of humans or in so far as they have value for us in other ways. The rationale to respect the robot here is not that the robot has moral agency or moral patiency, but instead that it belongs to a human and has value for that human person and that in order to respect that other *human* person we have certain indirect obligations towards robots-as-property. Historically, this reliance of property has also been one way to protect slaves—at least to protect them from violence by humans who are not their master—and it is conceivable that society will use similar arguments with regard to robots in the future. Since things are valuable to us (humans) for various reasons, it is likely that robots will receive some degree of indirect moral consideration anyway. This is plausible since some humans tend to value some things more than they value (other) humans. It is conceivable that some expensive intelligent robots will be regarded as so valuable, that harm will be done to human beings in order to protect these robots. Violence has always been one of the means humans use to protect things they see as their 'property' and things they value in other ways. Whether or not the institution of private property can be justified, it is a way of thinking about moral status we need to take into consideration when discussing the future of robotics and indeed the future of humanity.

Conclusions

These reflections on 'moral robots' can contribute to a better understanding of not only robot morality but also and especially of human morality. Dealing with the question of what kind of ethics we should build into robots challenges us to scrutinize the assumptions of our normative moral theories, our theories of emotions, and our theories of moral status. Should morality be rule-based? Are emotions necessary for moral reasoning? Is the ability to have mental states a necessary condition to have emotions? How much (certainty about) 'reality' does the social-moral life require?

I have argued that *if* we understand morality as requiring (among other things) the capacity to exercise emotions, and *if* having mental states or consciousness are necessary conditions for having emotions too, then the prospects for us ever being able to build 'moral' robots are dim. It is unlikely that in the foreseeable future there will be robots with real mental states or consciousness, and as long as this

[14] In animal ethics this demand for consistency is known as 'the argument from marginal cases'.

[15] Note that these moral categories constituted (and arguably still constitute) a kind of moral life that is fundamentally asymmetrical. An alternative, symmetrical moral framework would accommodate perceptions and treatment of robots as companions or co-workers. One might also apply other 'human' categories to them. However, I will not further discuss this issue here.

is the case, perhaps we had better refrain from trying to design highly intelligent, autonomous machines, since if such machines are rule-based but have no emotions, they will not be capable of engaging in genuine moral reasoning and hence they might even be dangerous. (Consider, for example, what highly intelligent military robots that lack any moral feeling could do to humans.) Of course this would not be an argument against building *different* kinds of robots that are much less autonomous and intelligent—they would not need to have moral agency or emotions; however, we would require moral agency and emotions from those who design, use and control these kinds of robots.

However, I have also argued that these demands on robot morality are too high, since in human morality we tend to rely on appearance. If intelligent autonomous robots were able to produce the appearance of being moral—including the appearance of emotions—and behave in ways that contribute to the moral life, we would have a good reason to be more optimistic about living with them (or at least to be as optimistic as we are about living with other humans). Thus, it would be unfair or at least inconsistent to require that robots must have real mental states, real consciousness, or real emotions in order to be moral.

Finally, I have also noted that to include human and non-human entities in our moral world, we use several conceptual tools that are content with appearance. If we consider ways of thinking about issues such as disability, slavery, and property, human morality turns out to be more appearance-based and more open to entities without mental states than standard moral theory allows. This challenges us to further reflect on the ways in which we regard and treat humans and non-humans—with and without mental states, consciousness, or emotions. This is not only helpful to robot designers but also to moral philosophers who rightly and understandably continue to be puzzled by the ways we use our moral capacities and how we define and justify the borders of our moral world.

Acknowledgments I wish to thank the reviewers for their pertinent questions and useful suggestions, which helped improve the paper's organization and fine-tune its arguments. I also thank Julie Bytheway for her advice on grammar and style and Nicole Vincent for copy-editing the final version of the manuscript.

References

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, 94–103.

Damasio, A. (1994). *Descartes' error: emotion, reason, and the human brain*. New York: G.P. Putnam's Sons.

De Sousa, R. (1987). *The rationality of emotion*. Cambridge, MA: MIT Press.

Foot, P. (2002). Hume on moral judgment. In *Virtues and vices*. Oxford/New York: Oxford University Press.

Goldie, P. (2000). *The emotions: a philosophical exploration*. Oxford: Oxford University Press.

Greene, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

James, W. (1884). What is an emotion? *Mind, 9*, 188–205.

Kennett, J. (2002). Autism, empathy and moral agency. *The Philosophical Quarterly, 52*(208), 340–357.

Merleau-Ponty, M. (1945). *Phénoménologie de la Perception*. Paris: Gallimard.

Nussbaum, M. C. (1990). *Love's knowledge*. Oxford: Oxford University Press.

Nussbaum, M. C. (1994). *The therapy of desire: theory and practice in hellenistic ethics*. Princeton: Princeton University Press.

Nussbaum, M. C. (1995). *Poetic justice: literary imagination and public life*. Boston: Beacon 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.

Solomon, R. (1980). Emotions and choice. In A. Rorty (Ed.), *Explaining emotions* (pp. 81–251). Los Angeles: University of California Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.

Wallach, W., & Allen, C. (2008). *Moral machines: teaching robots right from wrong*. Oxford: Oxford University Press.