sciendo

Kairos. Journal of Philosophy & Science 20, 2018
Center for the Philosophy of Sciences of Lisbon University

# Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering

Mark Coeckelbergh
University of Vienna and
De Montfort University (UK)
mark.coeckelbergh@univie.ac.at

**Abstract.** This paper tries to understand the phenomenon that humans are able to empathize with robots and the intuition that there might be something wrong with "abusing" robots by discussing the question regarding the moral standing of robots. After a review of some relevant work in empirical psychology and a discussion of the ethics of empathizing with robots, a philosophical argument concerning the moral standing of robots is made that questions distant and uncritical moral reasoning about entities' properties and that recommends first trying to understand the issue by means of philosophical and artistic work that shows how ethics is always relational and historical, and that highlights the importance of language and appearance in moral reasoning and moral psychology. It is concluded that attention to relationality and to verbal and non-verbal languages of suffering is key to understand the phenomenon under investigation, and that in robot ethics we need less certainty and more caution and patience when it comes to thinking about moral standing.

**Keywords:** moral standing, robots; empathy; relations, language, art, phenomenology, hermeneutics, Wittgenstein.

**DOI** 10.2478/kjps-2018-0007

## 1. Introduction

Many people respond to robots in a way that goes beyond thinking of the robot as a mere machine. They empathize with the robot, care about it. Consider for instance the case of a robot dog called "Spot" that was kicked by its designers.[1] The people of the company (Boston Dynamics) wanted to show that the robot had the capability to stabilize itself again. But this was not the only result. Some people reacted by saying things such as "Poor Spot!", "Seriously Boston Dynamics stop kicking those poor robots what did they ever do to you?", "Kicking a dog, even a robot dog, just seems wrong", "The robot remembers. This is why the robot uprising starts". Another case is HitchBOT,[2] a robot which hitchhiked and then was vandalized in Philadelphia: its head and arms were ripped off. Here too reactions were similar: "America should sit in the corner and think about what it's done to poor HitchBOT"[3] is just one of them. And the HitchBOT people had the robot tweet: "My trip must come to an end for now, but my love for humans will never fade. Thanks friends".

Given that robots are machines, these reactions are puzzling. How can we explain and understand them? How can we explain these people empathize with robots? Is this "abuse" of robots? And, morally speaking, *should* we empathize with robots? Is it wrong to kick a robot? How can we philosophically support the intuition that there might be something wrong with "abusing" robots?

This paper inquires into the phenomenon of empathy with robots in various ways. First it reviews some psychological experiments in order to further describe the phenomenon of empathy with robots. Then it discusses normative ethics: could there be any justification for the intuition some people have that there is something wrong with kicking or "abusing" robots? However, then the paper gradually leaves the psychological and the normative ethics discussion by asking the question regarding the moral standing of robots. After considering some standard

1 http://edition.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html

2 http://mir1.hitchbot.me/

3 http://paleofuture.gizmodo.com/hitchhiking-robot-lasts-just-two-weeks-in-us-because-hu-1721544551

answers, it is argued that the standard approaches to ethics of treatment of robots and to their moral standing are problematic in so far as they are too distant and uncritical. The paper explores how these problems could be remedied by pointing to a more relational way of thinking about ethics and moral standing, by highlighting the importance of language as a condition of possibility of moral standing ascription and normative ethical reasoning, and by further reflecting on appearance, using some artistic work. It will be concluded that such attention to relationality and to verbal and non-verbal languages of suffering is key to understand the phenomenon under investigation.

## 2. The Psychology of Empathy with Robots

There is interesting psychological work emerging on empathy with robots. Suzuki et al. (2015) used electroencephalography to investigate neural responses to robot "pain" and compare these to response to human pain. The researchers showed pictures of painful and non-painful situations involving human hands and robot hands, such as a finger that is going to be cut with a knife. The result was that although people empathized more with humans than with robots, in both cases there was what the scientists interpret as an empathic response. It seems that we empathize with humanoid robots. Perceived pain (or threat of pain) seems sufficient.

Human capability to empathize with robots is also confirmed in other experiments, for instance in work by Rosenthal et al. (2013) and by Kate Darling (2015). In a workshop with the Pleo robot,[4] for instance, Darling found that participants hesitated to torture robots. First participants were asked to name the robots and play with them. Once they were done, they were asked to torture and kill the machines. Most people hesitated. In an experiment with hexbugs, which are not even humanoid, similar results were observed. First people observed the robots for a while and then they were asked to smash them. Some were told a backstory that attributes lifelike qualities or the robot was given a personal name. Participants hesitated significantly more to strike the robot

4 http://www.bbc.com/future/story/20131127-would-you-murder-a-robot

when it was introduced through such framing. (Darling 2015). As Darling remarks elsewhere, we tend to treat robots different than toasters, even if they are not very advanced yet: "our robots are nowhere close to the intelligence and complexity of humans or animals, nor will they reach this stage in the near future. And yet, while it seems far-fetched for a robot's legal status to differ from that of a toaster, there is already a notable difference in how we interact with certain types of robotic objects." (Darling 2012).

How can we deal with these results from a philosophical perspective? One possible response is doing robot ethics, understood as the application and discussion of normative moral theories and as a discussion of moral standing.

## 3. Robot Ethics as Normative Moral Theory and as a Discussion about Moral Standing: What's Wrong with Torturing a Robot?

If at least some people have the intuition that there is something morally wrong with kicking or torturing a robot, even if it is supposed to be "just" a machine, a thing, how can this intuition be justified and supported by arguments?

Two major normative theories do not seem to provide any support. Traditionally, our deontological rules are about humans. According to this view, kicking or "torturing" a robot is not wrong since there is no moral rule or law against it, and no moral duty not to torture a robot. The act is not forbidden. Moreover, such acts are not wrong from a consequentialist point of view (e.g. utilitarianism) since in contrast to humans and animals a machine can feel no pain. No harm is done to robot, it feels no pain, there is no suffering, the robot is not hurt. Therefore, a consequentialist argument does not work either to support the intuition that there is something wrong with it.

A better chance has the Kantian argument about animals. Kant held that humans are "altogether different in rank and dignity from things, such as irrational animals, with which one may deal and dispose at one's discretion." (Kant 2012, 127) But nevertheless he argued that we have *indirect* duties towards animals:

> "So if a man has his dog shot, because it can no longer earn a living for him, he is by no means in breach of any duty to the dog, since the latter is incapable of judgment, but he thereby damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind … for a person who already displays such cruelty to animals is also no less hardened towards men." (Kant 1997, 212).

Now this argument can be applied to robots. As Darling puts it: "The Kantian philosophical argument for animal rights is that our actions towards non-humans reflect our morality — if we treat animals in inhumane ways, we become inhumane persons. This logically extends to the treatment of robotic companions." (Darling 2012).

The way Darling formulates the argument also reminds of what I think is probably the strongest ethical argument for not mistreating robots: the argument from virtue ethics. According to this argument, mistreating a robot is not wrong because of the robot, but because doing so repeatedly and habitually shapes one's moral character in the wrong kind of way. The problem, then, is not a violation of a moral duty or bad moral consequences for the robot; it is a problem of character and virtue. Mistreating the robot is a vice.

Such a virtue ethics approach may be an answer to the concerns Whitby (2008) raises about mistreatment of robots. Whitby rightly points out that we may learn from the discussion about computer games: the question whether violence in games lead to violence in real life: the worry is that 'they might do it for real', in other words, that the violence carries over to life outside the game context. But although Whitby sides with those who claim that video games lead to desensitization of users to violent activities in real life, the empirical support for this claim is controversial, to say the least. It is likely that this would also be the case for a similar discussion robots: someone "abusing" a robot may not abuse human beings. Still we might have the intuition that something wrong is going on when someone "tortures" a robot. Now virtue ethics may provide a way out of this problem: instead of using a consequentialist kind of reasoning, which would need empirical support for the claim that there

are empirical consequences of mistreating robots for the treatment of humans, and assumes strict and problematic distinctions between real and virtual and between "robot contexts" and "human contexts", we can say that there are bad consequences for the *moral character* of the person. The point is that the abusive behavior towards robots may or may not lead to bad consequences for treatment of other humans, but in any case has bad consequences for the person as moral subject.

However, there is still something problematic with these ways of reasoning about non-humans. They concern humans; they are not about the robots. The concern is with the virtue of the human – *regardless* of what happens to the robot. Is this total focus on the moral subject instead of the moral object justified? What about the moral standing of robots as moral patients, as entities on the receiving end of moral action? If they get any moral standing at all in the Kantian and virtue ethics approach, then it is a rather indirect or what we may call a "weak" form of moral standing: they get only moral standing indirectly via the moral standing of the human moral subject.

Indeed, one way of approaching the problem is not so much to use moral normative theory, which is usually focused on humans, but on the discussion about the *moral standing* of robots. Are they just machines, or are they more than machines, and what are the moral implications for their standing? Is there a good argument for giving robots direct or "strong" moral standing?

In this discussion, the "default" or "common sense" position denies that machines can ever have moral standing. No matter how automated and interactive they might be, robots are technologies used by humans, as Johnson says in the moral agency discussion (Johnson 2006). And if this is the case, one could infer, they do not have moral standing as patients (nor are they moral agents – but this is not our concern here). In the moral patiency discussion, Bryson argued that robots are property and that therefore we are not morally obligated by them (Bryson 2016). While Bryson leaves open the possibility that they may be moral patients in the future, for now they do not meet criteria for moral standing.

This result is unsurprising, since traditionally these criteria are modelled on humans: sentience, consciousness, having mental status, having

the ability to suffer, and so on. The bar is high, and too high for robots. In response, one could of course lower the bar. Using Foridi and Sanders' approach to moral agency (2004), for example, one could argue that moral patiency needs to be dealt with in a similar way: by moving away from criteria such as mental status and rather embracing a 'mind-less' and non-anthropocentric morality which holds that a sufficient degree of interactivity, autonomy, and adaptability warrants moral standing as moral patient. If, for a specific robot, we find a level of abstraction at which we can observes these properties, then according to this kind of approach we could ascribe moral standing to the robot. However, the problem with this response is that most of us have the intuition that having properties such as interactivity, autonomy, and adaptability is not enough to give the robot moral standing as patient. In spite of these properties, it could be argued, the robot remains a machine. It does not experience or feel anything if we "abuse" it.

This outcome is unsatisfactory. Does this mean that empathic responses to robots are bound to remain unexplained and unjustified? How can we philosophically address this problem in a way that is different from these answers?

A starting point could be to recognize that some robots may appear to humans, and may be treated by humans, *as if* they have feelings, consciousness etc. (Coeckelbergh 2010a). Can we do more to take seriously this phenomenon, philosophically? But this already starts changing the question, since it is no longer about the moral standing of the entity as such, but about what happens between human subject and non-human object. The following section draws on my previous work on moral standing (Coeckelbergh 2010a; 2010b; 2011a; 2011b; 2012; 2014) and refers to work by David Gunkel (2012; 2017) to deconstruct the very discussion about moral standing, about robots and other entities.

## 4. Questioning the Question and Deconstructing the Procedure: Towards a more Relational Approach

The way disagreements about moral standing are usually dealt with in moral philosophy is by focusing on the properties of the entity in question. If one follows this moral procedure and argumentation, one can

determine the moral standing of an entity. First the ontological status of the entity in question is determined, then it its moral standing is inferred from this ontological status. The reasoning goes as follows:

1. all entities of type A with properties P, Q, R, … have moral standing S
2. entity X is of type A (and has properties P, Q, R, …)
3. therefore, entity X has moral standing S.

For instance:

* all conscious entities have moral rights
* this entity is conscious
* therefore, this entity has moral rights.

For example, on this basis robots are denied moral rights: they are deemed to be not conscious. However, this reasoning raises some epistemological problems. With regard to the first premise, how can we be so sure that all entities of a particular type have a particular kind of moral standing? The history of morality shows that we have often been mistaken about this, for instance when we have failed to give any moral standing to animals, slaves, etc. It is not clear where we can find firm ground for this, unless we are dogmatic. With regard to the second premise, how can we be sure that a particular entity has the morally relevant property in question? Scientists always discover new facts about non-human entities, for instance about fish and about plants, and since we cannot "look into the head" of an animal or really know what it means to experience the world as that particular animal, it is not clear on what basis we can make firm conclusions.

Furthermore, there is a deeper problem with the procedure as a whole: it presupposes and creates distance. As I argued in *Growing Moral Relations* (2012), this moral procedure puts entities at a distance, performing a kind of moral dissection of its ontological and morally relevant properties. The entity is reduced, so to speak stripped down, to its relevant moral properties. We look at the entity from a god's eye point of view, from what Nagel called a 'view from nowhere'. What is missing is the concrete relation to the entity in question, for instance an animal or indeed a robot, and indeed the entity as a whole when it is encoun-

tered in a concrete situation, context, and narrative. A more *relational* approach is needed.

The approach I proposed, then, takes seriously the phenomenology and experience of other entities such as robots, and sees moral standing not as the starting point but rather as the *outcome*: moral standing is itself the outcome of the process of relation and interaction. Discussions in robot ethics often assume an outdated, Cartesian metaphysics and epistemology, which starts from objects and subjects as given and fixed. Influenced by the phenomenological tradition, I propose to see subject and object as mutually interdependent and mutually constituting. We need a 'relational' approach then in the sense of an epistemology that takes seriously this relation between subject and object. Moreover, we also need as *social*-relational approach (Coeckelbergh 2010b; 2014), in the sense that the robot may appear as a quasi-other; this turns the question about "status" into a question about social relations: should I or we have social relations with the robot? Moreover, and this is perhaps the more fundamental philosophical point since it enables a more critical approach: when I, as a moral subject, "ascribe" moral status to an entity, I am not the first one to do so and the way I do it and the status I ascribe are probably already available in my society, my culture, and my language – more generally in what Wittgenstein would call my 'form of life' (Coeckelbergh 2017). Therefore, the question of moral standing is always connected to the question who is part of the moral *community* and what moral games are already played when and before I ask the question. We need to reveal and criticize the social background of the question.

One way of further developing this point is to use transcendental language (Coeckelbergh 2012; 2017). If we consider how we ascribe moral standing, then one could say that the usual kind of reasoning about "moral standing" is blind to the *conditions of possibility* of that ascription. Once we take into account an epistemology where subject and object are no longer independent from one another, we have to consider how this ascription comes about and what must be presupposed on the side of the subject – which then shapes the object. In particular, it seems that the appearance of the entity is very important (since we cannot know what the entity in question really thinks, feels, etc.) and that there is already the personal experience and narrative of the moral

subject which is linked to a social and cultural background, including particular uses of language, that shapes and makes possible a particular ascription. One could say that there is already a kind of "moral grammar" when and before I "ascribe", which shapes my ascription of the moral standing of an entity. Moreover, the subject is also shaped by the object: the way we deal with other entities, the way we experience them, what we say about them, the way we treat them, and so on, also *says a lot about me and says a lot about us*. If we take a truly relational approach, it means that when we are putting other entities in question, we are also putting ourselves – as persons and as humanity – in question. This is somewhat disturbing, perhaps, but it is a logical consequences of realizing that we are *already* related before we ask the question, and that the way we ask the question is itself not morally neutral. In other words, what is missing in the standard discussion about "moral standing" is a more critical approach.

Let me make this more concrete for the case of robots. In the experiments of Darling, for instance, how people spoke about and to the robot mattered a lot for how people treated the robot. (See also Coeckelbergh 2011b). This shows that, whatever one may think of how people *should* treat the robot, descriptively the moral standing of the robot depends on how people talk about it and to it. Indeed, language is an important condition of possibility of our use of technologies (Coeckelbergh 2012; 2017). There is no such thing as a robot-in-itself or thing-in-itself. In practice, it seems, there is no such thing as an independent object isolated from the subject; the subject also shapes the object and vice versa. The robot question or, to use Gunkel's words, the machine question (Gunkel 2012), puts also *us* in question. Moreover, through narratives people start caring about the robot; there is also a relation in the sense of "relationship" – even if only one-sided. If we take into account this moral psychology and moral epistemology, then, we arrive at a philosophical account of moral standing ascription that is in line with the findings of the empirical work and that, in contrast to normative ethical theories or discussions about moral standing, is not only about *either* humans (traditional ethical theory) *or* about robots, but about how humans and robots relate to one another. This account is not distance since it attends to how people in practice treat the robot and it is critical since it does

not just talk about the moral standing of an isolated and abstract technological object "robot" but about a moral-epistemic and moral-psychological relation and process in which that object is constructed and practically related to, leading sometimes to caring about the robot or to empathy with the robot. It enables us to be critical about the moral reasoning itself, which presupposes a particular language and narrative, for instance a scientific one that constructs the object as "machine" or a care narrative in which the robot becomes a pet or a companion. Finally, it is also a more historical approach, which is sensitive to various ways in which we have categorized robots and machines in the past, for instance as slaves – which in turn relates to a problematic social history of human-human relations.

This relational approach to moral standing is compatible with, and supported by, David Gunkel's interesting efforts to think 'otherwise' about machines (Gunkel 2012; Coeckelbergh and Gunkel 2014; Gunkel 2017). Gunkel has used Levinas to connect the discussion about moral standing with that about alterity. According to Levinas, we should not start from ontology and then go to ethics, but exactly the other way around: ethics precedes ontology. The other comes first, and we are first obliged to respond to the other. The question, then, is no longer what the moral standing of the entity is, or indeed what the entity "is" at all, but rather how to respond. First there is the encounter with the situated other; this is the starting point, not our self (Gunkel 2017). Robots, then, are (potential?) others, to which/whom I respond. Again, language is important here: language is used to make a difference between the "who" we include in our moral community and "what" remains excluded. Again, words are not morally neutral (Coeckelbergh and Gunkel 2014).

For normative ethics (how should we relate to robots, which moral standing ought we ascribe to robots), this relational and more other-oriented approach means at the very least that we should be less certain about the moral standing of any entity, including those entities we call robots. While from a distant so-called "objective" point of view we can call robots mere machines and ascribe the standing of a thing to them, once we consider how humans relate to and talk about robots in various ways, we come to see that this "objective" categorization is one particular way of relating to robots and by no means the only one. Moreover, it is a

particularly distant way of relating to robots and it is one that is uncritical about how this "objective" approach is itself made possible and shaped by language use – e.g. the language of science. Therefore, instead of immediately fixing the moral standing of robots, it is recommendable to be more careful when making decisions about this and to be critical of how this ascription comes into being – that is, to be critical of the conditions of possibility of that ascription, which means to be critical of the way we talk about machines, animals, and so on. This does not mean that in practice one should no longer make decisions about moral standing – at some point we may have to make such decisions, also as communities and societies – but it invites us to take a more patient and cautious attitude and reflect more on how we make such ascriptions. We should no longer do it without asking questions about how subject and object are entangled in various ways; without considering this phenomenology and hermeneutics of moral standing ascription, our ascriptions are too distant and too uncritical.

Further work in this area may be needed to engage more with philosophy of language, for instance the work of Searle, which helps us to understand how objects are given social meaning through particular uses of language (declarations and status functions), or the work of Wittgenstein, which shows how use of language is always embedded in language game and a form of life. Recently I have reflected more on what Wittgenstein's insights could mean for thinking about technology, which included a response to Searle (Coeckelbergh 2017). Let me offer the following suggestions for how their work on language may be used for better understanding the moral standing of robots.

Searle (1995) argued that we give social meaning to objects by using language, in particular by so-called 'status functions'. For instance, a bank note only counts as money if we collectively declare it to be money. Similarly, one could say that the moral standing of robots is a 'status' that is socially declared: it is a physical object which is given meaning by means of language. Hence we could say that in the case of empathy with robots there is a tension between, one the one hand, a socially accepted meaning of robots as mere machines and mere things (one supposes here that there is a collective declaration to this effect), and on the other hand a moral-psychological experience that reveals robots as more than

machines – a meaning which may become part of the social but perhaps not yet. Language use is still focused on robots as things. Moreover, Searle maintains a strict distinction between the material object (e.g. the piece of paper) and the meaning we ascribe to it. But is this strict distinction tenable? If the relational view outline above is adequate, then it seems that subject and object are much closer related. This would support a stronger version of constructivism.

Another view which helps us to understand the relation between language and robots, which may support such a strong version, and which is particularly helpful to explain why it is so difficult to change our language about technology, is Wittgenstein's view in the *Philosophical Investigations* (1953). For Wittgenstein, the meaning of things depends on the use of language, and this use in turn depends on what he calls 'language games' and a 'form of life'. Wittgenstein's point was that meaning is holistic and varies with use; there is not one fixed meaning attached to an object. Moreover, in contrast to what Searle's view may suggest, for Wittgenstein meaning is linked to an entire form of life; we cannot simply change the rules or change that form of life. For thinking about robots, this view implies that the way we see and experience robots is shaped by games and a form of life in which robots are talked about and treated as things, and that it may be very difficult to change this since there are already rules, there is already a "grammar" that frames this technology in certain ways. Of course we may try to change the game. And by designing different technologies, these may function as a kind of game changers. But the changes will generally be small, more like the banks of the river slowly change (to use a Wittgensteinian metaphor).

Maybe both Searle's and Wittgenstein's view support my point that to say that language is a condition of possibility for moral status ascription does not imply that we can simply change our language use as individuals in order to give robots a different moral standing (understood as status). Rather, the way we individually relate to robots – through language – will always be shaped and made possible by a kind of social contract (Searle) or a form of life (Wittgenstein) which already pre-frames robots in certain ways, and which may or may not be in line with our normative moral theories or with concrete experiences we have when we encounter robots (experiences such as empathy).

## 5. The Role of Art

To open up our thinking about robots and invite further critical hermeneutical work, it may also help to engage with artistic work. For instance, Kris Verdonck has created installations in which machines seem to be in distress[5] or are "tortured" or "killed".[6] In each case the appearance of suffering is created, not by verbal means but my movements and sounds of the machine. On the one hand, we know that it is a machine. On the other hand, we may feel an ethical, possibly empathic response. In the relation between subject and object (epistemologically and metaphysically speaking), and in the concrete confrontation with the machine, something happens which creates this empathy. We recognize and respond to the language of suffering. At the same time we lack a socially accepted language to talk about our experience since we see "things". There is no collective language available to talk about more-than-things (or not yet? no longer?), and in our language games and our form of life we tend to strictly separate between humans and non-humans, humans and things, humans and machines, etc. Like in the experiments mentioned in the beginning of this paper, we experience a discrepancy between our language/thinking and our perception/feelings.

In an installation by Bill Vorn and Louis-Philippe Demers, called La Cour des Miracles,[7] machines seem to be in pain and groan. Again there is the appearance of suffering, the display of what one may call "fake" suffering. The machine has been collectively declared to be a mere thing. We are used to play the game of sorting humans and non-humans, putting them in the "right" kind of ontological category. But at the same time, we may feel a real empathic response, confronted with these machines. The artists write:

> "By creating this universe of faked realities loaded with "pain" and "groan", the aim of this work is to induce empathy of the viewer towards these "characters" which are

5 http://www.atwodogscompany.org/en/projects/item/158-dancer-1?bckp=1

6 http://www.atwodogscompany.org/en/projects/item/164-dancer-2?bckp=1

7 http://woodstreetgalleries.org/portfolio-view/la-cour-des-miracles-bill-vorn-and-louis-philippe-demers/

> solely articulated metallic structures. Therefore, we want to underline the strength of the simulacra by the perversion of the perception of these animats, which are neither animals nor humans, carried through the inevitable instinct of anthropomorphism and projection of internal sensations, a reflex triggered by any manifestation that challenges our senses." (*ibid.*, footnote 7).

Again we learn something about humans and our relation to others – "fake" or not. Again we are revealed as the kind of beings who are able to feel and respond to the language and appearance of suffering, as relational and empathic beings. This language and this communication seems *also* part of our form of life – with the emphasis on "life": we are not only cultural but also embodied beings (and hence when we interpret Wittgenstein we should not only pay attention to rules but also to embodied knowledge; however, I will not further develop this here.) Moreover, works of art such as these invite us to destabilize and critically question established meanings and borders, here to question the sharp border between machines and humans, or at least invite us to consider how in our imagination and feeling we already easily cross this border – whatever science or metaphysics may tell us. Even if our traditional metaphysical frameworks forbid these crossovers, as social, embodied, and relational beings we create empathic bridges to non-humans, including machines. Apparently our form of life is open enough to make these bridges, to allow for these cracks, frictions and uncomfortable tensions.

Of course one may object that the suffering is "fake". But the work of art also invites us to reflect on this: how sure are we in the case of humans and animals that the suffering is real or fake? And is it not better to err on the side of caution when ascribing moral standing, given that epistemologically speaking we can never be so certain about other entities and given that we know now that there are conditions of possibility such as language in play when we ascribe that moral standing? And is the language game of moral standing ascription itself not problematic, since we already take distance from the "entity", rather than responding directly and immediately, possibly in empathic ways and in ways that may help the entity – just in case the suffering is real? Of course I also "know" that

these are machines. But at the same time I do not have certainty and I may have other knowledge and know-how: knowing someone else's suffering and knowing how to help. Instead of sticking with only one type of morally relevant knowing, we might first have to explore different kinds of knowing and experiences and negotiate these various kinds of knowing. We first may have to respond. Perhaps we *should* do so to be on the side of caution. Such a cautious, patient, and open attitude (and indeed character), then, can be said to constitute a meta-moral demand and a meta-virtue or moral-epistemic virtue.

## 6. Conclusion

In this paper I started from the phenomenon that people seem to empathize with machines and have the intuition that there is something morally problematic about "abusing" robots. In order to better understand what is happening in such cases, I have taken a number of different approaches and turns. I have reviewed work in empirical psychology, I have offered some arguments from normative ethics, and I have discussed moral standing. The latter has enabled me to introduce and further develop some arguments that question distant and uncritical moral reasoning, and add some more philosophical work to the empirical results. In particular, I have argued for a more relational, engaged, and cautious approach to the question, which attends to how language frames our thinking about the question and how in concrete confrontations with other entities appearances of suffering play an important role – possibly in tension with accepted collective meanings and mainstream forms of life. More work – in philosophy and in art –is needed to further reflect on communications of suffering and our response to suffering, and on the role of language in these processes.

To conclude, in order to understand the phenomenon of empathy with machines, in order to understand why sometimes some people really seem to care about machines and their "suffering", and in order to conceptualize the tensions and frictions but also the bridges, it seems we need approaches that take seriously human subjectivity and human sociality as mediated by various kinds of languages and as shaped by human culture and human embodiment. To work on this and further investi-

gate languages and communications of suffering, different approaches are needed, not just empirical psychology or abstract moral reasoning focused on properties of entities or on application of moral theories. For ethics of robotics and similar types of ethics, my stress on the meta-virtues of caution and openness invites us to be less certain and to explore more – possibly with the help of artistic work. In this way, we can further learn about robots and especially about human beings, who are able to empathize and to respond, and who care.

## References

Bryson, Joanna. 2010. Robots Should Be Slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). Amsterdam: John Benjamins.

Coeckelbergh, Mark. 2010a. Moral Appearances: Emotions, Robots, and Human Morality. *Ethics and Information Technology* 12(3): 235–241.

Coeckelbergh, Mark. 2010b. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology* 12(3): 209–221.

Coeckelbergh, Mark. 2011a. Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *Philosophy & Technology* 24(3): 269–278.

Coeckelbergh, Mark. 2011b. You, Robot: On the Linguistic Construction of Artificial Others. *AI & Society* 26(1): 61–69.

Coeckelbergh, Mark. 2012. *Growing Moral Relations: Critique of Moral Status Ascription*. Basingstoke and New York: Palgrave Macmillan.

Coeckelbergh, Mark. 2014. The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology* 27(1): 61–77.

Coeckelbergh, Mark. 2017. *Using Words and Things: Language and Philosophy of Technology*. New York and Abingdon: Routledge.

Coeckelbergh, Mark and Gunkel, David. 2014. Facing Animals: A Relational, Other-Oriented Approach to Moral Standing. *Journal of Agricultural and Environmental Ethics* 27(5): 715–733.

Darling, Kate. 2012. Extending Legal Protection to Social Robots. IEEE Spectrum 10 Sept 2012. Retrieved 22 June 2017 from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots.

Darling, Kate. 2017. 'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. *Robot Ethics S 2.0*, eds. P. Lin, G. Bekey, K. Abney, R. Jenkins, Oxford University Press.

Floridi, Luciano and Sanders, J.W. 2004. On the Morality of Artificial Agents. *Minds and Machines* 14(3): 349–379.

Gunkel, David. 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. Cambridge, MA: MIT Press.

Gunkel, David. 2017. The Other Question: Can and Should Robots Have Rights? Ethics and Information Technology (online first). DOI 10.1007/s10676-017-9442-4.

Johnson, Deborah G. 2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4): 195–204.

Kant, Immanuel. 1997. *Lectures on Ethics*, eds. P. Heath and J.B. Schneewind. Trans. P. Heath. Cambridge: Cambridge University Press.

Kant, Immanuel. 2012. *Lectures on Anthropology*, eds. A.W. Wood and R.B. Louden. Cambridge: Cambridge University Press.

Rosenthal-von der Pütten, Astrid M., Krämer, Nicole C., Hoffmann, Laura, Sobieray, Sabrina, and Sabrina C. Eimler. 2013. An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*. 5: 17–34.

Searle, John R. 1995. *The Construction of Social Reality*. London: The Penguin Press.

Suzuki et al. 2015. Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography. Nature, Scientific Reports 5.

Whitby, Blay. 2008. Sometimes It's Hard to Be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents. *Interaction with Computers* 20: 338–341.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*, eds and trans: Hacker PMS, Schulte J. Oxford: Wiley-Blackwell, 2009.