Philos. Technol. (2014) 27:61–77
DOI 10.1007/s13347-013-0133-8

RESEARCH ARTICLE

The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics

Mark Coeckelbergh

Received: 16 July 2013 / Accepted: 11 September 2013 / Published online: 20 October 2013

Abstract Should we give moral standing to machines? In this paper, I explore the implications of a relational approach to moral standing for thinking about machines, in particular autonomous, intelligent robots. I show how my version of this approach, which focuses on moral relations and on the conditions of possibility of moral status ascription, provides a way to take critical distance from what I call the "standard" approach to thinking about moral status and moral standing, which is based on properties. It does not only overcome epistemological problems with the standard approach, but can also explain how we think about, experience, and act towards machines—including the gap that sometimes occurs between reasoning and experience. I also articulate the non-Cartesian orientation of my "relational" research program and specify the way it contributes to a different paradigm in thinking about moral standing and moral knowledge.

Keywords Moral standing · Moral status · Moral relations · Moral knowledge · Robots · Machines · Descartes · Levinas: modernity · Moral change · Phenomenology

1 Introduction

In fields such as the military and space exploration, but also in health care and households, humans are increasingly confronted with robots. Humans are assisted by robots and interact with robots. Some of these robots become increasingly independent of (direct) human intervention and plans are made to increase their autonomy and intelligence. For example, in the military field, research aims at making drones that can operate without a human "in the loop". And in healthcare, there are "companion" robots for the elderly that act like pets.

The idea of developing autonomous intelligent robots challenges philosophers to ask the question concerning moral standing again. Are these robots "mere machines" or do they deserve moral standing? Are they things or should they be added to the list

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

of previously excluded others such as slaves, foreigners, women, and animals? Should moral agency be reserved for humans, or can these machines be regarded as moral agents, perhaps having a "machine morality"? Can autonomous intelligent machines be considered as tools, or are they moral agents and/or moral patients? Is their moral status akin to, say, a hammer, or are there similarities with the status of animals or even humans?

Common sense tends to go with the former view. At least, if we are *asked* if robots have moral status—that is, if robots "in general" have moral status—most of us tend to answer negatively. We do not generally *think* of robots as having moral status and we would have a hard time *arguing* that they have any kind of moral agency or patiency. We strongly hesitate to declare "robot rights" or ascribe any other kind of moral status to the robot. We *think* of them as "mere machines". At the same time, however, people who interact with robots (and, sometimes, computers) often do this in a way that suggests otherwise. They ascribe emotions and intentions to the robot (good or bad intentions), take care not to "hurt" the machine, or even love it and care about it.

For example, people who play with robotic pets like the dinosaur-looking Pleo or elderly people who use the baby seal looking care robot Paro tend to treat the robot like a pet or a child: they take care the robot does not fall down, they stroke and cuddle it, talk to it in ways people talk to babies or pets, and so on. In other words, they treat the robot not as a "mere thing" or a "mere machine", but as "more" than that. More generally, humans tend to anthropomorphize, or at least zoomorphize: we treat the robot as if it has human or animal properties—including moral status. Is this simply irrational? How can we explain this response, and how can we make sense of this gap between what we *think* about robots and how we *act* towards them?

In this paper, I will first describe what I take to be the "standard" approach to thinking about moral standing. I will then argue that this approach has a serious "internal" and "external" problem: it has epistemological difficulties and it is unable to make sense of the mentioned gap between thinking and doing. In response to these problems, I draw on previous work (Author 2012a) to offer a different, relational approach and show that it can overcome these difficulties by opening up the possibility of a non-Cartesian moral epistemology and by contributing to a radically different way of thinking about moral standing—one which Gunkel has recently described as a "paradigm shift in moral thinking" (Gunkel 2013).

2 The Standard Approach: The Machine and its Properties

2.1 Standard Reasoning About Moral Standing

Discussions about moral standing usually assume that moral standing depends on properties. For example, in animal ethics the "classical" debate between Regan and Singer is about what property is morally relevant: is it "being subject of a life" (Regan 1983) or is it, following Bentham (1879), the capacity to suffer (Singer 1975)? While both have distinct approaches when it comes to normative moral theory (deontology versus utilitarianism), both assume that in order to determine moral standing, one should investigate if the entity in question has the morally relevant property.

Similarly, with regard to robots one can argue that the relevant property should be consciousness or the ability to suffer, for example, or one can propose different criteria and properties (see for example the debate on the moral agency of artificial agents in Floridi and Sanders (2004), Himma (2007), and Sullins (2006)). But again in these discussions of moral agency and patiency, the assumption is that moral standing depends on the entity having a particular property.

Put in terms of moral status, the form of reasoning is:

entity *x* has property *p*
any entity that has property *p*, has moral status *s*
entity *x* has moral status *s*

This is the way Western culture has reasoned in order to "emancipate" slaves, women, and (partly and some) animals. Moral science has investigated these entities in its philosophical–scientific laboratories. With its Platonic inheritance, it has stripped away the phenomena and has revealed the presence of the relevant property (or properties): for some entities, it turned out that "yes, they can talk", that "yes, they can suffer", etc. and they received moral status as moral agents, moral patients, or both. Other entities, including many "members of the animal kingdom", remain excluded from the communities of moral agents and moral patients.

2.2 Problems

But this approach raises at least two epistemological problems. Consider the first premise of the argument. How can we know that a particular entity *x* really has a particular property *p*? Properties such as "being subject of a life", "capacity to suffer", or "consciousness" are notoriously difficult to determine and claims are often contested. Skepticism tells us that we can never be certain about the internal states of other entities. We cannot directly observe properties such as "being subject of a life". Therefore, it seems difficult to establish the first premise.

The second premise is equally problematic. How do we know for sure that a particular property *p* justifies moral status s? Do we have access to a moral metaphysics, a Book of Values, in which we can find propositions about moral status that cannot be doubted? And how *can* we be so sure in the case of new entities such as autonomous intelligent robots? Again, a skeptic response seems in order here. Of course we do not know for sure, we might be wrong about our presuppositions. And if both premises can be doubted, the conclusion is also doubtful. We cannot know for sure that a particular entity *x* has moral status *s*.

Moreover, even if these epistemological problems could be solved and we could reach certainty about the moral status of a particular entity, this does not solve the problem of the gap between reasoning and experience, between thinking and action, between belief and feeling. Even if there were a moral metaphysics, or indeed a moral *science*, which could tell us the truth about the moral standing of a particular entity, we may experience the entity differently when we interact with it and feel differently about it. For example, we might feel that some intelligent and autonomous machines are "more than machines" or we might address some of them with "he", "she", or even "you". Consider again people who have robot pets: usually, these people know very well that their robot is just that: a robot, a machine. Yet their experience and

behavior is different. Consider also how we experience robots that look like humans. We might sense their "presence" when they are in the room. We might feel that they are "companions" if they "live" in our home.

How should we respond to this gap between beliefs and behavior, between reasoning and experience? The moral–scientific answer to this problem is then that we are simply incorrect about the entity's status. There *is* a gap, so the answer goes, but there *should not* be a gap. But this answer does not help us to understand how we act towards these robots and it makes all responses except the "correct" one appear as irrational. We cannot make sense of experiences that depart from the idea that a robot is a machine and we have to dismiss them as "childish" or "ignorant". We have to say: "Don't you know this is a machine?" But is this the only possible answer? Is it the *best* answer?

In order to overcome these two problems—the epistemological one and the one about the gap between reasoning and experience—it is not sufficient to make amendments to, say, the Other Minds problem or to put the label of anthropomorphism (or personification) on particular experiences. Instead, I propose to start challenging the question. Instead of asking about properties we could ask about relations: relations between entities, and also the relation between (human) *subject* of moral status ascription and the object of moral status ascription. Let me explain this approach and discuss its implications for the discussion about machines as moral agents and patients.

3 An Alternative, Relational Approach: Moral Relations and the Conditions of Possibility of Moral Status Ascription

3.1 A Relational Approach

In *Growing Moral Relations* (Coeckelbergh 2012a), I have argued for a relational approach to moral status, which sees moral status as something that emerges through relations between entities. We can find this view in ecophilosophy, for instance, which considers natural relations, or in Marxism, which considers social relations. The idea is that an entity cannot be defined without reference to its relations—both social and natural relations (and more precisely, also mixtures of these). For example, a particular animal has its place in the ecosystem and in the webs of social relations with other animals. These relations also have a history and are tied to specific places, habits, and things. To define moral standing in isolation from these relations is itself a moral violation, since it takes as its departure an abstract "entity" with "properties". Even the very term "animal" constitutes already such an abstraction and hence violation.

Applied to robots, this approach would mean that in order to determine their moral standing, one would need to know its relations with other machines and with humans. We need to know the situatedness, history, and place of the machine. We need to know how it is naturally, materially, socially, and culturally embedded and constituted. We need to contextualize moral standing, rather than study the entity as an atomistic curiosum in the anatomic theater of moral status science.

The view that moral standing depends on relations has normative appeal: relationality and relational thinking seem something good. It also seems to fit well

with ecological currents in environmental thinking and with ethics of care. However, the danger of replacing properties by relations is that the approach becomes dogmatic and that it only inverts the view, from properties to relations. Then "relations" become yet another abstraction. Then, the entity is part of a "moral ontology" and is abstracted and violated once again. Then we get again a moral metaphysics or a moral science—this time a relational one—but this does not solve the problems identified, since how can we be sure about the relations of entities and their moral qualities, and what if our experience of an entity or indeed of a relation is not in concord with what moral science or metaphysics tells us? What if we do not experience the animal or the robot in terms of "relations" and "relata"? What if our common moral language and experience does not fit the metaphysics?

Therefore, we need to take a next step. We should not only (and not so much) consider relations between entities, but also (or instead) the relation between the subject and object of moral status ascription. Moral standing is not an "objective" property "out there" which we can define without discussion. Discussions about moral standing take place within language and thinking, more specifically within human language. Humans are the subject of moral status ascription. This implies that the status of the object of moral standing is not independent from human subjectivity. The entity appears to us in a particular way, and its appearance depends on human subjectivity, that is, on human thinking and also on specific thinking and specific cultures and forms of life at a particular point in history. This means that with regard to moral standing we should not only (and not so much) discuss ontology (e.g., properties and/or relations) but also (and rather) *epistemology*. Which knowledge and what kind of knowledge can we have of entities and of their moral standing? What kind of *epistemic* relation emerges between subject and object, and by which (other) relations is this epistemic relation shaped? Let me further explain and develop this point.

The properties view assumed that an entity has only one "correct" ontological status and meaning, which is contrasted to the "appearance" and "perception" of the robot. Those who accuse people who behave otherwise than they "should" rely on the science of moral standing which assumes a dichotomy between the entity—e.g., the robot—as a *Ding-an-sich* and the appearance of the entity. But we can think of an alternative, nondualist epistemology that rejects this dichotomy and accepts that there are several ways in which an entity can appear to us, with none of these ways of seeing having a priori ontological or hermeneutical priority. Some ways of seeing may be better than others, but this evaluation has to take place with regard to particular entities, practices and experiential situations, can allow of various perspectives, and cannot be pre-*determined* "before" by a metaphysical properties ontology. Abstract reasoning can be part of this hermeneutics, but it alone cannot provide an "end-solution" that solves the problem of moral standing once and for all.

I have argued that the moral status of an entity is constructed and, to the extent that we cannot control it, grows, on the soil of relations we have with an entity as epistemic subjects. This approach brings in human subjectivity. Relational thinking then does not, and should not, constitute a new ontology, a better, "correct" view of the world, but should focus on relationality between subject and object, on how we "world", on how we relate, epistemically and morally. This means that philosophy has to keep open the discussion rather than try to close it, in order to prevent a closure

that is itself an ethical violation. It also means that a philosophical discourse on moral standing does not only focus on direct arguments about moral standing, but also questions how moral standing comes into being, how moral patients are constructed. This "meta-ethics" can then ensure that when philosophers (or anyone else) makes direct arguments about the moral standing of particular entities, there is a position from which to criticize discourses and practices of moral status ascription themselves—including philosophical discourse.

For robotics, this moral phenomenology means that we can acknowledge that robots can appear in different ways to different people in different situations and contexts, and that the status "mere machine" is not necessarily and not automatically the "correct", only, or best way of constructing the entity, but that this appearance and construction is itself "in question". What matters, morally speaking, is how the entity appears (Coeckelbergh 2012a, p. 24). Thus, our ethical attention is shifted from ontology to epistemology, from object to subject, from "what things really are" to how *we* look at things. This enables us to be aware, for instance, of historical changes in the construction of machines, animals, women, slaves, etc. and of cultural differences in these matters and therefore to critically reflect on our own and on today's constructions of "machines", "companions", "artificial slaves", etc.

This does not mean that we can no longer say that some constructions and ascriptions of moral status are better than others. But the process of evaluation involves a broader hermeneutics than one that relies on science and philosophical criteria/principles alone. Interestingly, it also opens the door for a different kind of intersubjectivity (one might say a true kind of intersubjectivity) than that of the standard approach. The standard approach presupposes that there is one truth about the entity's status and moral standing, and a philosopher working in that tradition (or anyone else, for that matter) will try to convince others of her view. This means that, in this view, there really is only *one* legitimate moral subject: one that is convinced by reason and science. In this tradition, to be a moral subject means to subject oneself to *the* moral truth.

The alternative view I attempt to articulate, however, allows of multisubjectivity and a plurality of truths. Moreover, this view also allows us to introduce subjects that are not merely "rational agents" but real people who are confronted with the entity, interact with the entity, do things to the entity (or not), and are faced with a problem that is not "the problem of moral standing" but a very concrete problem of how to relate to particular entities. For instance, when one faces a human-looking robot, one might be unsure about how to behave towards it. In other words, this approach introduces the moral subject to whom the moral standing of particular entities matters. This subject does not subject itself to reason but has to subject itself to the reality of the concrete experience and reality of the entity-in-relation, the entity which is here and now and confronts me, "asks" me. Moral standing is then not an abstract philosophical question but the practical question how to relate and how to respond.

Thus, this relational and phenomenological approach also makes room for focusing on the morality of the concrete human–robot relation, interpreted as entangled with subjectivity. We can describe such a relation from the outside, from what Nagel called a "view from nowhere", but we can also attend to the phenomenology of experiencing the robot. A properties approach singles out one particular type of experience, one that is produced in the lab and in the study room. But the problem

of moral standing does not (only) arise there. It arises in the first place in hospitals, drone control rooms, homes of people—all kinds of places where robots may play a role. It is in these particular places and particular situations that the question needs to be answered: How to relate to this entity?

3.2 Using Levinas and Haraway: the Face of the Robot and our Entanglements with It

As Gunkel has suggested, my view can be interpreted as "a Levinasian gesture" given my insistence that relations are prior and given Levinas' claim that ethics, not ontology, is prior (Gunkel 2013; Levinas 1969). Levinas' approach emphasized the ethical relation with the other—and, within this context, the significance of the "face" of the other. It is in the concrete confrontation with the other, who appeals to my responsibility, that the question of moral standing arises. Then ethical evaluation is not something that takes place after some kind of scientific or metaphysical examination of an entity, which distances us from the entity and already predefines its status *as* an "entity", but instead is the starting point. If I meet a human being or an animal, then the "face" may ask me to respond, may appeal to my responsibility.

This connection to Levinas, together with Gunkel's own work on what he calls "The Machine Question" (Gunkel 2012), may inspire us to ask the question *whether machines can have a face*. Can they appear as, and can they be constructed as, an "other" at all? And if not, can some machines *become* others if our relations to them change, or rather: if *my* relation to *one* of them changes when I encounter it (or him, or her)? The answer to this question cannot be given by reference to particular properties, but needs to be discovered in the living phenomenology of daily experience, and indeed needs to involve those who are left out in the standard approach: the humans who "meet" the robots, work with them, interact with them. For example, we need to involve the experience of people who take care of robot pets in their home and elderly people who have a robotic companion. Thus, in this kind of thinking, there is room for taking seriously concrete moral experience, for making sense of unique encounters between humans and machines, and for discussing their moral significance.

However, there are at least two crucial ways, we must depart from Levinas. First, as Gunkel also notes Levinas reserved morality to humans and restricted the moral relation to a human–human relation, whereas the approach I offer here opens up the possibility of thinking about other kinds of relations and their moral quality: relations with animals, with robots, etc. I do not make a straightforward normative claim such as "Robots should be considered as others" but rather argue that whether or not they appear as others, whether or not they have a "face", is precisely what is at stake in the discussion about moral standing. Thus, the notion of moral standing needs to be related to interpretations of encounters and interactions between humans and machines. Ascribing moral status to machines (or not) is part of the ways we make sense of these encounters and interactions, and luckily moral experience is in no way limited to "ascribing moral status" or "discussing moral standing" of particular entities. We need a richer moral hermeneutics that reveals the very term "machines" as already normatively loaded and that learns from concrete human–robot encounters (real or imaginary) in order to take seriously human moral subjectivity-in-action and in-relation.

For this purpose, we can learn from Haraway's work, which meditates on encounters between humans and animals. In her book *When Species Meet* (2008), she explores human–animal encounters and questions the idea of human exceptionalism, the idea that there is a "Great Divide" (Haraway 2008, p. 12) between humans and nonhuman entities, which pervades Western thinking and which unnecessarily and regrettably limits the range of interpretative possibilities we have with regard to relations with nonhumans. Haraway's approach is not to engage in abstract thinking alone, but also to discuss concrete stories and concrete human–animal relations. A similar book could be written about concrete relations with robots. Attending to concrete human–robot relations is likely to reveal a more rich palette of moral and other meanings than the one offered to us by moral science. For our moral thinking to make progress, we need stories of encounters with, say, a robot dog, at least as much as we need more abstract arguments about "moral standing" (and indeed reflections such as the one offered in this paper).

Guided by Gunkel's scholarship, it is instructive to take a closer look at Haraway's comments on Derrida (Gunkel 2012, pp. 122–123). In *The Animal That Therefore I Am* (Derrida 2008) reflects on his cat and wonders how the cat looks at him and if the cat "responds" to him. On the one hand, Derrida is concerned with a concrete, unique cat—*his* small female cat. Derrida writes: "the cat I am talking about is a real cat (…). It isn't the *figure* of a cat. It doesn't silently enter the bedroom as an allegory for all the cats on the earth". The cat that looks at him is "*this* cat I am talking about" (Derrida 2008, p. 6). On the other hand, as Haraway says, side-tracked by "his textual canon of Western philosophy and literature" (Haraway 2008, p. 20) Derrida is concerned about his nudity and does not consider different forms of engagement, in particular *mutual* engagement. As a man in the bathroom, he knew much more about this, for example he knew he was in the presence of the cat, but as a *philosopher*, he "had no idea" about the meaning of the "bodily postures and visual entanglements" involved here (Haraway 2008, p. 22). Haraway writes:

> I am prepared to believe that he did know how to greet this cat and began each morning in that mutually responsive and polite dance, but if so, that embodied mindful encounter did not motivate his philosophy in public. That is a pity. (Haraway 2008, p. 23)

Similarly, one may meditate on concrete interactions between human and robots—not only robots in general or a robot as an allegory for all robots on the earth, but on interaction with *this* robot, for example on the "presence" of a human looking robot (e.g., in the context of an encounter with a particular humanoid robot) and on the question if there is a sense in which humans can "dance" with robots (e.g., in the context of a relation between a particular person and a particular robotic pet).

Here too, we might be side-tracked by the "textual canon of Western philosophy and literature" when we describe these interactions and encounters and get lost in words that are alienated from the phenomenology of the encounter. Due to our philosophical and scientific tradition, discussions about moral standing and, more generally, about "robots" tend to disregard embodied encounters and engagement with robots, and focus on "the" robot and its properties, often defining it a priori as a "machine". (I will say more about our philosophical tradition in the next section of this paper.) This narrows down the range of possible experiences and interpretations.

For example, if we—as philosophers or scientists—immediately "correct" ourselves when we see a human looking robot and talk about "the robot", "it", "the machine", etc., then we close off other experiential possibilities. Perhaps we first had a very different experience and impression, for example we might have felt a kind of "presence" or we might have felt that the robot was "watching" us.

Even Gunkel's Levinasian approach, which might inspire us to discuss the "face" of the robot (see also Coeckelbergh 2012b), risks to get lost in logo-centric accounts of *the* face of *the* robot if it does not open up to a different kind of body of knowledge: knowledge we do not have as (Western) philosophers, but as beings that engage in concrete interactions and indeed "dances" with other entities. A phenomenology of human–robot relations needs to rely on embodied knowledge and knowledge-in-relation—even if this means breaking through the public/private barrier and overcoming our shame.

Second, in contrast to Levinas, one should not only consider I–you relations. My book also shows that moral status ascription depends on many conditions of possibility that have to do with relations that go beyond the I–you relation: linguistic relations, social relations, technological relations, spiritual relations, and spatial relations. How a particular robot appears to *me* (or how I construct it), and indeed how my concrete relation with that robot is shaped, does depend on how *we* talk about robots (e.g., "machines" versus "companions"), on how *we* humans live together and live with robots, on the technological developments in *our* society, on *our* culture and its religious dimension which encourages certain status ascriptions rather than others, etc.

It is what I called, with a Wittgensteinian term, our "form of life" that makes possible, but also constrains how we relate to robots. Our moral experience is "personal", concerns the "I", but at the same time related to how "we" perceive and evaluate entities, including animals, machines, other humans. Our personal construction of the robot is influenced by the way our culture constructs machines, and this construction is not only a word process but also a living process, it emerges from a living and changing whole we call "society" and "culture". There are already patterns of interpretation, patterns of action (habits), patterns of living, and indeed patterns of evaluation. There are already norms and values. When we talk and write about moral standing, we do not start from a blank slate.

Derrida's shame is part of that form of life, and when we, as philosophers, "meet" a robot, we also carry with us the weight of the same philosophical tradition. Looking into the mirror of robots, we may feel ugly, imperfect, and impotent. And like Derrida, at some point, we might feel the "gaze" of the robot, we might feel that the robot "looks" at us. More generally, robots may confront us with our own vulnerability: the contrast between our vulnerable bodies and the seemingly invulnerable machine may be uncomfortable. We are ashamed, and want to hide our vulnerability. An artist like Stellarc, by contrast, is not afraid to face that confrontation and shows his aging body together with the powerful, perfect machine (see for example his performance "Host Body Coupled Gestures" in the early 1990s). Can Western philosophers overcome their vulnerability phobia? Can we overcome what Anders called our "promethean shame" (Anders 1956, pp. 23–25)?

Inspired by Haraway's work, we may wonder if robots will remain "machines" or if they can become companions. Will people start saying, as they tend to say of

people who have "met their dog" (see also Haraway 2008, p. 301), that someone has "met her robot"? Would such a person, having *that* kind of relation with *that* robot, still feel shame at all in front of the robot? And is there, at that point of personal engagement, still a *need* to talk about the "moral standing" of the robot? Is not moral quality already implied in the very relation that has emerged here? For example, if an elderly person is already very attached to her Paro robot and regards it as a pet or a baby, then what needs to be discussed is that relation, rather than the "moral standing" of the robot.

Of course it may be that true companionship will never be possible with machines. One could argue that because machines are not vulnerable, or at least not vulnerable in the same way as we are, they cannot become real companions. One could argue that *they cannot eat the same bread* (I play here with the etymological meaning of "companion", which has also been noted by Haraway: eating the same bread, *pan*) and that they cannot *eat* and *digest* the world in the same way as we do, that is, they have a different epistemic relation to their environment, a different openness to the world—indeed, perhaps no true openness at all. But in any case, we need to open up more epistemic and hermeneutic possibilities than the ones given by the "machine" vocabulary in order to make sense of experiences, some of which are *already* interpreted in terms of "companionship".

3.3 How the Relational Approach Overcomes Problems with the Standard Approach

For the purpose of this paper, it is now important to understand that and how the relational approach I have outlined here overcomes the two problems identified in the beginning of my paper.

First, the epistemological problem is overcome since the question "What do you know about that entity?" is rendered obsolete. If we no longer assume two atomistic, unrelated entities but a relation, there is no longer a need to bridge an epistemic gap, since if there is a relation the gap is already bridged. The answer to the problem is: We *already* know the entity as and since we are *already* standing in relation. There is already a relation *before* we talk and think about the question of moral standing or moral status. At the moment that we think about moral standing of the robot, a linguistic, social, material, etc. relational structure is already in place, has already grown, which for example make us think of the robot as a "machine" or an "it" rather than a "companion" or a "he". There is already a relation: a concrete, I–robot relation but also other relations that make possible, constrain, and shape that concrete relation. For example, if there is a relation between a particular elderly person and a robot, there are also other relations, for example in the nursing home, that influence the human–robot relation. For instance, other people might encourage the relation, condemn it, and so on.

The knowledge that arises from these relations—including knowledge about the robot—is usually "tacit" knowledge (to use Polanyi's term); usually, it is not made explicit. Unfortunately, the standard approach relies mainly on explicit, propositional knowledge retrieved from the sciences and from moral philosophy, understood as a moral science. The moral philosopher then "knows", for example, that a particular entity is not conscious and concludes that it cannot have moral standing. But this disregards any knowledge that stems from nonscientific areas of experience,

including the experience of the people who relate in a far more direct way to the entity. If we, philosophers, want to say something about the moral standing of machines, why not talk to the people who design these machines and use them in their daily life? I do not see any reason to discount their knowledge a priori, and the fact that their knowledge is (also) tacit is certainly not a good reason to exclude that knowledge from exercises in moral hermeneutics.

Second, with this approach there is no longer a gap between, on the one hand, a "correct" way of seeing the robot and my "perception" of the robot. Instead of the "moral status" of the robot, what matters is the moral quality of the relation and if there is a need to talk about how to morally consider the robot at all, its moral standing is something that grows within the moral relation; it is not prior to it. At the same time, based on the framework offered by my analysis of the conditions of possibility of moral status ascription, we can explain the gap between thinking and doing, between reasoning and experience as originating in the relational structure of our current form of life. I suggest that in so far as our culture is a *modern* culture, it does not "allow" us to see robots as more than things, more than machines: the way we talk about them (at least, *as philosophers* and *as scientists*), the way we define sociality (excluding things), the way our culture categorizes the world, etc. does make it difficult to relate to robots in a different way than we do now.

I argue in my book that the only way our moral thinking can change is if these (more structural) relations change: if these moral relations grow into different ones. If specific relations grow such that we start talking differently about particular kinds of robots, include them in our social world, and if our technologies and our societies develop in a way that facilitates this, etc., then we might think and reason differently about them as well. Partly this is happening, but the dominant paradigm is still modern: there is a divide between we (humans) and machines (things), and morality—including moral standing—is only to be found on the side of the humans.

Let me further inquire into this relation between the problem of moral standing of machines and the problem of Western modernity by discussing what we might call our *Cartesian* legacy.

4 The Possibility of a Non-Cartesian Moral Epistemology: Toward a Different Paradigm

4.1 Questioning Cartesian Thinking About Humans and Machines

For understanding the dominant moral paradigm, and indeed for understanding the problems identified, it is helpful to refer to Descartes' work, which is usually seen as a crucial and central articulation of modern thinking. Moreover, reading Descartes is especially instructive for understanding our modern view of "machines" and their relation to humans, and is therefore highly relevant to the present inquiry.

First, when it comes to the problem of not being able to make sense of moral experiences that go beyond seeing robots as "mere machines", we should refer to Descartes' philosophical construction of an ontological divide between humans and especially the soul of humans, on the one hand, and the human body, animals and machines, on the other hand, which he made in his *Discourse on Method* (Descartes

1637). This divide has contributed to a dualistic way of thinking about the relation between humans and machines, which renders it difficult to think otherwise about them. It has become so much a part of our form of thinking and our form of life that we have difficulties making sense of experiences that do not easily fall within these conceptual categories.

More, as Gunkel has insightfully argued, this philosophical operation has rendered the machine "not just one kind of excluded other [but] the very mechanism of the exclusion of the other" (Gunkel 2012, p. 128). It means that we exclude machines from moral standing, but also that we use the very concept of the "machine" as a way to exclude robots, (some kinds of) animals, etc. We argue that "because they are machines" they are not worthy of moral standing. In that sense, "the machine question" is not just a "marginal" discussion about the moral status of autonomous intelligent robots and (other) artificial agents, a discussion that could be ignored by moral philosophy, but is central to understanding the modern Western paradigm of moral thinking. It shows that modern ethics has always been a "machine ethics".

This Cartesian mechanism of exclusion is also illustrative of a more general current of "negative anthropology" that runs through the history of Western culture, which has always defined humans in terms of what they are not: we have been defined as non-gods, non-animals, non-machines, and indeed as non-beast-machines or at least more-than beast-machines. Today, we may be defined as non-robots or more-than-robots. Apparently in the West, we have always needed and used other entities in order to *distinguish* ourselves as humans. For this purpose, we prefer entities that are sufficiently similar to us but not—so we emphasize—the same. For example, when it comes to animals, we insist that we are not apes, not pigs, not dogs, not cats, etc. When it comes to robots, we also choose robots that look similar to humans: androids, humanoid robots. We use them as hermeneutic devices to construct and maintain the Cartesian divide and what Harraway calls the Great Divide. Anything that crosses this divide, or attempts to do so, is experienced as embarrassing if not threatening. Robots are used as purification tools: instruments that keep open the divide, that guard the borders between human and nonhuman.

Second, the epistemological problem also has its origin in modern, Cartesian thinking. In my description of the problem, I assumed that doubt necessarily has to take a kind of absolute form, that we should strive for absolute certainty. The argument asks us to be in the position of the Cartesian doubter, who doubts everything except his own thinking. The "Other Minds" problem, whether applied to humans or to robots, can be seen as an application of this Cartesian procedure: it is doubted if we can know anything at all about other minds (natural or artificial). But why should we strive for absolute certainty? As I suggested before, we already have knowledge of other entities if and when we relate to them.

Moreover, as Torrance has rightly remarked when commenting on my book at AISB in Birmingham (Torrance 2012), the problem definition is itself Cartesian and arguably belongs to the "standard" or "dominant" way of thinking I criticize. Thus, one could say that I have beaten the standard approach at its own game. This is fine, but we cannot stop there. If I say that we can have *some* knowledge of other entities without a need for absolute certainty, then I have to articulate what knowledge and what kind of knowledge *is* available (rather than no knowledge at all) about other entities. In this paper, I have suggested that this is a more tacit kind of knowledge,

which emerges within the lived experience of the relation with the entities in question—here the concrete engagement with autonomous intelligent machines. The question is not "What can we know about the other entity?" We already know. We have knowledge-in-relation.

4.2 Beyond Cartesian Thinking

A different moral paradigm, then, must be non-Cartesian in these two senses: it should (1) overcome its obsession with keeping up distinctions between humans and machines as the main goal of thinking about moral standing in order to open up space for articulating different kinds of (moral) experience which should not be discounted a priori and thus (2) turn to the phenomenology of concrete engagement with entities in order to acknowledge an epistemic range and an action repertoire that otherwise does not appear on the moral radar screen: a spectrum of human experiential and relational possibilities that has been suppressed by modern categorization and Cartesian doubt. When it comes to morality, we should be especially sensitive to those kinds of experience and relational situations when such doubt is simply out of place, when categorization is not only unnecessary but itself an act of violence, when the only "reason" or "ethical imperative" is one that is already there in the situation and in the relation.

It remains to be seen if relations with robots can make this kind of ethical relation possible. Personally, I doubt it (based on my experience with current robots and on my speculations about future robots). But even today there might be room for moral experiences between the extremes of "mere machines" and Levinasian "others", and it is important to realize that the very way we talk about robots (e.g., in terms of machines, talking about "it", etc.) is not neutral but already implies a moral stance, and that this moral stance is made possible by larger patterns in our culture that encourage certain ways of interpretation and action rather than others.

To handle *that* moral challenge is quite a different task than examining the properties of an entity. It changes the game of moral standing in such a way that what is at stake is not only the particular entity in question but also what the human is and *my* position as a moral subject. Questioning the moral standing of other entities is also asking: Where do *I* stand? This proximity may be uncomfortable for us modern thinkers, since the approach demands for us that we acknowledge our moral involvement with other entities and our moral complicity in the world, which we can no longer look at from a distance or imagine that it does not exist. (Consider for example the myriad ways in which our lives are entangled with animals, for example by means of our eating practice. We cannot deny the relations. We are complicit. We are "forced" to question our own position.)

For thinking about the moral standing of autonomous intelligent robots, this non-Cartesian orientation means that we have to give up the moral–philosophical and moral–scientific project of determining their moral status in order to reach moral certainty understood as a normative framework built on the foundations of a moral-metaphysical order. Instead, we embark on a different philosophical journey, which attends to the moral challenges posed by concrete human–robot relations and which reflects on what may lie beyond the boundaries of modern ethical thinking.

In order to explore this kind of thinking, we might want to learn from non-modern elements in other, non-Western cultures. It could also be helpful to look at

the non-modern elements in our own culture, in "pagan" culture but also in our contemporary technological culture. For example, if we can find traces of animistic experience in that culture, then such an interpretation can contribute to thinking otherwise. In any case, the project of a new paradigm in moral thinking should not be understood in modern terms, that is, as being about something entirely and absolutely *new*. Experience is always historical. And like moral status, moral change is subject to many conditions of possibility. It depends on change in our language, in our way of living together, and so on. Changing moral thinking means changing a form of life—including its linguistic, social, and technological dimensions.

Robots are already part of our form of life, and this should be taken into account if we embrace the project of paradigmatic change in moral thinking. If I am right about the conditions of possibility of moral change, then whether or not such a project succeeds is itself dependent on how our relations with robots will change. It is difficult to predict what will happen when robots will become more autonomous and intelligent. Our technological culture and our technological practices might grow into forms we cannot foresee yet. We might try to steer things, of course, but we should also be aware that the power of the word, of *logos*, is limited. Whether we like it or not, the moral outlook of world and the role machines will play in it depends not only on words but also on flesh, fluids, wires, and silicon.

This is difficult to accept for those who think *text* is prior and who try to turn everything else into *context*. Even postmoderns, who claim that context is prior, remain within this logos-oriented thinking. The relational paradigm I wish to contribute to, by contrast, starts from relations and does not limit its scope to linguistic relations and philosophical gestures. The "radical interpretation" Derrida talks about (Derrida 2008, p. 160) can only succeed if the new meanings grow in those wider relational ecologies, and not outside of it—if they emerge from concrete relations and not if they become disconnected from experience. If we should be ashamed of anything at all, it is not our nakedness in front of others, but our epistemological and moral egocentrism and solipsism, that deprives us of an openness and a curiosity that would enable us to be sensitive to the appearance of others and to truly *respond* to others—to *their* nakedness, to their vulnerability.

This openness also implies that we acknowledge our own vulnerability as moral subjects. Within a concrete relation there is doubt, of course, but *that* kind of doubt is of an embodied kind: it *shows* how we look, in our posture, in our gestures, in the sound of our voice. It shows in the way we interact with our environment, in the way we actively relate to it. It is not the doubt of Descartes, who used thinking to attempt to delete his embodiment and his worldly existence. It is the doubt that is felt, the doubt that can hurt, the doubt that *matters*.

I do not know if intelligent autonomous robots can and will ever appear to us as such true others, and there are other, more urgent concerns (e.g., about animals), but at least here we have a new moral question. The question is no longer "Does it think" or "Does it suffer?"; the new question concerns the otherness of the robot and the question *where we stand* and what the "we" is. The doubt is no longer about the properties of the robot, but directly about the moral questions how to relate to the robot: Is this robot part of "us"?

Whether or not philosophers ask this question is not the main point of concern; the point is that the question *asks itself*, or not, in concrete relations and within our technological culture. Philosophers can contribute to interpretation, but the word should not be confused with the sound and the breath on which it depends. Logos does not come first. Moral standing can neither be declared nor decided, as Gunkel suggests. Moral thinking is not so much a matter of "deciding who or what is to be included" in the "we" (Gunkel 2012), but to at most to influence, to some extent, the process of the formation of these moral boundaries.

Perhaps the extent to which we can influence this process is more limited than philosophers assume. The game of thinking about moral standing is itself dependent on the dynamics of concrete relations and on the continuously living, exchanging, and evolving cultural whole those relations are part of and to which they contribute. Who or what is inside or outside the moral community, who or what we can morally relate to, cannot simply be declared or agreed upon. Who or what is part of the "us" cannot and should not be forced but has to grow.

4.3 Relations with Other Humans

Note that this question regarding the moral boundaries of our moral community does not only concern relations with machines or animals, but also relations between humans. The relational and non-Cartesian approach articulated can and must also be applied to the question regarding the moral standing of humans. What does the "game change" proposed here imply for how we should relate other humans? Some humans do not currently enjoy the status of (full) "moral agent" and there are situations in which some humans are not even treated as full "moral patients". How can the relational approach proposed here help us to better understand what is going on in these cases and how can it contribute to moral improvement?

In my book, I have commented on the implications for the "human" question and in this article I have mentioned the history of oppression and emancipation (e.g., of "slaves", of women), but further work is needed to spell out the more concrete implications for our practices, to distinguish it from existing approaches to the standing of humans (e.g., human rights), and to compare it to other approaches that attend to relations (e.g., ethics of care, Marxist thinking, etc.).

5 Conclusion

Should we give moral standing to machines? In this paper, I have argued that an approach that focuses on moral relations and on the conditions of possibilities of moral status ascription, provides a way to take critical distance from what I call the "standard" approach to thinking about moral status and moral standing, which is based on properties. The different way of approaching moral standing proposed here does not only have normative appeal and overcomes epistemological problems with the standard approach, but can also explain how we think about, experience, and act towards such robots—including the gap between reasoning and experience. Furthermore, comments on my work by Torrance and Gunkel have inspired and encouraged me to articulate the non-Cartesian orientation of my relational research program and

to better specify the way it contributes to a different paradigm in thinking about moral standing and moral knowledge.

The result is a proposal for a moral hermeneutics that relies on a wider area of experience and knowledge than the standard approach and that has implications for how we approach other entities and for how we relate to ourselves as moral subjects. Thus, although this paper does not make direct normative philosophical arguments about moral standing—it does not make claims about the moral status of autonomous intelligent robots, for instance—it instead shows that such arguments are part of a larger approach to moral thinking and indeed *a way of doing things* in philosophy and science, a moral paradigm. This enables us to take distance of the standard approach and, by doing that, to bridge the distance with the entities. I have argued that the standard approach unnecessarily and undesirably divides up moral reality between humans and nonhumans, and that this way of doing things is not morally neutral, but unnecessarily and undesirably reduce the variety of interpretations and of possibilities to relate to other entities—to the point of making possible violence. Justifying moral standing by referring to properties of entities is not morally neutral since, in taking distance, in exercising a detached attitude, it also makes it possible to exclude others.

Maybe we cannot avoid the game of inclusion and exclusion, but where to draw the line should not be presented as being a matter of which properties an entity has. The alternative paradigm I have tried to articulate enables us to ask different questions, which are not about the properties of machines and other entities, but instead question whether we should call them "machines" at all and whether "they" are part of "us" or not. It also asks us to engage far more closely with the entities and those that are directly related to them. Ultimately, it asks us to give up the illusion of invulnerable moral subjectivity. We need to take a stance. Finally, I have explored the question what role thinking plays in the history and future of growing moral relations between humans and other entities, and what kind of moral knowledge we need to cope with the challenges these relations present. I have suggested that with regard to this moral coping, the love of wisdom should not be equated with the love of words.

References

Anders. (1956). *Die Antiquiertheit des Menschen (volume I): Über die Seele im Zeitalter der zweiten industriellen Revolution* (p. 1987). München: C.H. Beck.

Bentham, J. 1879. An introduction to the principles of morals and legislation. J. H. Burns and H. L. Hart (eds). Oxford: Oxford University Press, 2005.

Coeckelbergh, M. (2012a). *Growing moral relations: critique of moral status ascription*. Basingstoke: Macmillan.

Coeckelbergh, M. (2012b). *Review of David J. Gunkel: The machine question: critical perspectives on AI, robots, and ethics*. Cambridge, MA: MIT Press.

Derrida, J. (2008). *The animal that therefore I am*. New York: Fordham University Press. Trans. D. Wills.

Descartes, R. (1637 (1998)). *Discourse on method*. Indianapolis, IN: Hackett. Trans. D.A. Cress.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gunkel, D. (2012). *The machine question: critical perspectives on AI, robots, and ethics*. Cambridge, MA: MIT Press.

Gunkel, D. 2013. Review of Mark Coeckelbergh's growing moral relations: critique of moral status ascription. Ethics and Information Technology (published online 28 Feb 2013)

Haraway, D. J. (2008). *When species meet*. Minneapolis, MN: University of Minnesota Press.

Himma, K.E. 2007. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? In: L. Hinman, P. Brey, L. Floridi, F. Grodzinsky, L. Introna. Enschede: Center for Telematics and Information Technology. pp. 163–180

Levinas, E. (1969). *Totality and infinity.* Pittsburgh, PA: Duquesne University Press. Trans. A. Lingis.

Regan, T. (1983). *The case for animal rights*. Berkeley: The University of California Press.

Singer, P. (1975). *Animal liberation*. New York: Random House.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*, 23–30. Retrieved from http://www.i-r-i-e.net.

Torrance, S. 2012. The centrality of machine consciousness to machine ethics. Paper presented at the symposium 'The machine question: AI, ethics, and moral responsibility', AISB/IACAP world congress 2012—Alan Turing 2012, 4 July 2012.