Minds and Machines (2021) 31:337–360
https://doi.org/10.1007/s11023-020-09554-3

GENERAL ARTICLE

Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking About Animals and Humans

Mark Coeckelbergh[1]

Received: 9 May 2020 / Accepted: 2 December 2020 / Published online: 30 December 2020

Abstract
The use of autonomous and intelligent personal social robots raises questions concerning their moral standing. Moving away from the discussion about direct moral standing and exploring the normative implications of a relational approach to moral standing, this paper offers four arguments that justify giving indirect moral standing to robots under specific conditions based on some of the ways humans—as social, feeling, playing, and doubting beings—relate to them. The analogy of "the Kantian dog" is used to assist reasoning about this. The paper also discusses the implications of this approach for thinking about the moral standing of animals and humans, showing why, when, and how an indirect approach can also be helpful in these fields, and using Levinas and Dewey as sources of inspiration to discuss some challenges raised by this approach.

Keywords Moral standing · Robots · Social robots · Indirect moral standing · Animal rights · Human rights · Relational approach · Levinas · Dewey

1 Introduction: Anthropomorphization of Personal Social Robots and the Discussion About Direct Moral Standing

The development of intelligent autonomous personal social robots has raised questions regarding their moral standing as 'moral patients' (Floridi 2013, pp. 135–136): if we consider what could be done to them, do we owe them anything? Can they be morally wronged? For instance, would it be wrong to "mistreat" them? Moreover, is it good, ethically speaking, that users treat them as more-than-things? Robots that act like animals, for example, may appear to be alive and conscious. Users tend to

✉ Mark Coeckelbergh
mark.coeckelbergh@univie.ac.at

[1] Department of Philosophy, University of Vienna, Vienna, Austria

treat them as pets. Social robots are explicitly developed to be "social" and to give such impressions. As Breazeal puts it: such a robot is designed such that 'interacting with it is like interacting with another person' (Breazeal 2002, p. 1). In contrast to, say, a teddy bear, users do not only project and play while knowing that it is not really an animal and a companion; here the autonomous and intelligent technology really co-creates the illusion. We do not yet know the full ethical effects of, say, robotic Teddy Bear 2.0, but it is clear that through the highly interactive features, users get much easier deceived into thinking that there is something real going on. Is this morally problematic, and if so, why?

The phenomenon of anthropomorphization of digital technology is well-known has been studied since decades. In the 1990s empirical studies based on conceptualizing the computer as a social actor (CASA) documented that computers are sometimes treated as if they are people (Reeves and Nass 1996); these computers thus already produced a perceived gap between what a technological artefact "is" and how it comes to be experienced in specific interactive situations. Robots have an even higher degree of 'anthropomorphizability', as is well-recognized in the social robotics and human–robot interaction communities (for an overview see for example Złotowski et al. 2015). Anthropocmorphism is one of the standard factors that is measured for social robots (Bartneck et al. 2009) and for example Turkle (2011) and Scheutz (2012) have identified ethical problems with anthropomorphization of robots. But smart assistants with voice interface also raise this issue: if we can now have conversations with machines, we might perceive them as persons rather than as things.

Linked to anthropomorphization is the phenomenon that people also tend to react empathically and protectively when they see a robot being "abused" or "tortured". For instance, as reports in the media show, some people reacted shocked when developers kicked a robot[1] or when a hitchhike robot was vandalized.[2] There is also empirical research that shows that people empathize with robots. For example, when told a story that personalized the robots, people hesitated to smash robots (Darling 2017). And Suzuki et al. show that people empathize with robots, even if they empathize less with robots than with humans (Suzuki et al. 2015). Whether or not these are cases of deception, here users also tend to treat social robots in ways that correspond to how we perceive and treat other humans, opening a similar gap between what the technology is supposed to be (a mere thing, a mere machine) and how users perceive it and interact with it. This raises normative questions as to whether robots that invite anthropomorphization should be built at all.

One way to approach these phenomena and responses from a philosophical point of view is to raise the question about the moral standing of such personal social robots. Can social robots be ascribed moral standing at all, and if so, why: how can this moral standing be justified? More precisely: Can social robots be moral patients

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

[2] https://www.cbc.ca/news/canada/hamilton/headlines/hitchbot-vandalized-in-philadelphia-fans-outraged-worldwide-1.3177193.

that deserve our moral consideration, and if so, why? This paper responds to what it identifies as two gaps or problems in the current thinking about this question:

First, during the past decade, the discussion about the moral standing of robots has been centered on the question concerning *direct* moral standing and the intrinsic properties of robots. This is the usual way philosophers proceed when reasoning about moral standing: they look at the intrinsic properties of an entity and then derive its moral standing. For example, in animal ethics it is typically argued that if an animal can experience suffering, then it has moral standing (consider for instance the work of Singer 1975). Based on this kind of reasoning, it seems wrong to give moral standing to robots, since robots do not seem to have any of the intrinsic properties necessary for direct moral standing. Any attempt to even *consider* robot rights, as for instance Gunkel (2018a, b) does when he presents different types of arguments concerning the issue of robot rights, has therefore met fierce opposition. For example, Bryson (2010) has argued that robots are property, and that we are not obligated by them. It has also been suggested that we could give robots direct moral standing based on lower-level properties of artificial agents such as autonomy and interactivity, transferring for instance Floridi's (Floridi and Sanders 2004) or Sullin's (2006) arguments for moral agency to moral patiency. But this is also counter-intuitive to many people; they tend to use anthropocentric criteria and emphasize the divide between humans and robots. According to Johnson (2006), robots are tools, no matter how interactive they are. However, this leaves a wide gap between, on the one hand, the experience people have of dealing with personal social robots, and, on the other hand, our normative moral concepts and reasoning. Is there no way this gap can be closed?

Addressing this gap, Coeckelbergh and Gunkel have developed a critical and relational approach to the issue of moral standing (Coeckelbergh 2010, 2012, 2014; Gunkel 2012, 2018a), which helps to understand why and how some people ascribe moral standing to robots, and how philosophically problematic our ways of ascribing moral standing are. The starting point is skepticism about what Coeckelbergh calls the 'properties' approach: to define (direct) moral standing based on properties of the entity. For example, it may well be that robots have no consciousness, but how do we really know? How sure are we about the intrinsic properties of humans? How can we be sure that an entity has particular intrinsic properties at all given that we cannot look into someone's mind and cannot be sure about the internal states of other entities (Coeckelbergh 2012, p. 14, 2014, p. 63)? As Gunkel (2018a) has noted, there is also a long-standing philosophical discussion how to define these properties, for example consciousness. This skepticism has given rise to the development of a more relational and critical approach (Coeckelbergh 2012), which focuses on how we relate to entities rather than on their intrinsic properties, and asks the critical question how moral standing is constructed, for instance by means of language and in social relations. With regard to robots, Coeckelbergh (2014) has argued for a social-relational approach that does not ascribe moral standing to machines on the basis of their properties but focuses on how that standing is shaped in our relationships with robots; it is not prior to it (71). For example, as we interact with robots, the language we use to talk about them and to them influences the relationship and the moral standing we ascribe to the robot. Also in a relational vein, Gunkel has

proposed an other-oriented approach (Gunkel 2012, 2018a, b), which draws on Levinas, Derrida, and others to change the standard properties-based question of moral standing to one that considers and discusses the inclusion, face, and perhaps even rights of machine others.

How does this approach help to address the gap I mentioned? A relational approach helps to conceptualize how it can happen, for example, that users care about a teddy bear or a personal social robot: users use language that constructs the bear or the robot as a kind of person, and they build up a "relationship" with it, which potentially constructs the robot as an 'other'. However, it is unclear what follows from such an approach for more practical and normative issues. At first sight, the fact that someone cares about her teddy bear or treats her robot as a person or other does not seem to justify its moral standing. At a descriptive level and as a matter of understanding the phenomena, a relational approach seem to work; together with scientific work on these matters, it gives us some insight into the phenomenology of human–robot interaction. But what is the answer to the *normative* question about moral standing? Do we owe anything to robots at all? Again there seems to be a gap between the descriptive and the normative. The question regarding the normative implications of a relational approach to moral standing needs more work than it has received so far in the literature.

Moving away from the discussion about direct moral standing and exploring some more practical and normative consequences of a relational approach, this paper addresses these gaps and offers four arguments why robots can be given what I will call "indirect moral standing" based on their extrinsic value for us, or more precisely, based on the way we relate to them. By discussing its normative consequences, the paper thus aims to further develop the relational approach and its implications for normative ethics.

To assist the reasoning, this paper uses analogies with what we may call "the Kantian dog" argument. It also uses the terms "direct" and "indirect", which Kant applies to duties, more broadly to make a distinction between arguments for direct moral standing, which rely on the intrinsic properties of the entity in question, and arguments for indirect moral standing, which give moral standing only indirectly in the sense that the argument is based on the moral standing of the entity who ascribes and gives moral standing (usually the human agent or subject), rather than the object and receiver of moral standing.

What is "the Kantian dog" argument? The challenge, formulated in line with Kant's example is the following: why should we avoid cruelty to a non-human entity if it is supposed to have no moral standing on its own? In his *Lectures on Ethics*, Kant famously argued that we have nevertheless 'indirect' duties to the dog. Kant formulates moral standing from the point of view of the human moral agent or subject: humans have only 'indirect duties' towards it—because of the potential implications for human character and human–human cruelty:

> '"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, p. 212)

Applying the terms direct and indirect more broadly, we can call this an argument for "indirect" moral standing since it ascribes moral standing to an entity via the moral standing of the agent who ascribes the standing (here the human), rather than the moral standing of the entity in question. The reason why we should not shoot that dog, according to Kant, has strictly speaking nothing to do with the dog as such but with *our* duties and virtue as humans.

Now such an indirect argument for moral standing can be used for thinking about robots. Robotics researcher Kate Darling already helpfully suggested that its logic extends to robotic companions: if we treat robots in "inhumane" ways, we become inhumane persons (Darling 2012). And earlier Coeckelbergh already alluded to this in his paper on moral standing (Coeckelbergh 2010). I will discuss this argument and use it in further arguments. I will also refer back to the teddy bear example, because it helps to distinguish between intelligent social robots and other things. However, the main aim of this paper is not to say something about dogs or teddy bears as such, or to discuss Kant's view of indirect duties. The teddy bear example and Kant's view of dogs are merely starting points for the present discussion. I will expand these suggestions into a more systematic and general framework that presents several reasons for giving indirect moral standing to robots, a framework which I then propose can be extended to the indirect moral standing of other entities.

Starting with robots, I propose that we should give moral standing to a robot if one or more of the following conditions is fulfilled (I say "if" but not "only if", since these are sufficient but not necessary conditions; there could be other reasons for giving them moral standing and even other reasons for giving them indirect standing):

1. If a human who would do something bad to the robot would be seen by other humans as having a bad personality, as being a bad person (in virtue ethics language: not virtuous, or as being in danger of doing bad things to humans;
2. If the human user has a (one-directional) relationship to the robot and has developed feelings of attachment and empathy towards the robot;
3. If the robot is part of a human–robot join action and collaboration and it is desirable or necessary that this collaboration continues;
4. If there is serious doubt about the robot's moral standing on the part of the user(s).

The rationale for choosing these conditions rather than others is my intuition that in the cases and phenomena mentioned above *humans* matter morally (whoever or whatever else matters morally) and that our thinking about moral standing should somehow do justice to *the ways we humans, as social, feeling, playing, and doubting beings, relate to robots*. In terms of philosophical resources used to elaborate these arguments and in order to illustrate the intuition that humans matter morally regardless of whoever or whatever else may matter morally, the starting point (but by all means not the end point) of this paper is Kant, in particular the mentioned argument about a man and his dog. Just as Kant assumed that the dog he talks about

has no direct moral standing, for the purposes of these arguments let us assume that the robots in question are "Kantian dogs", that is, that they are believed not to have direct moral standing or at least that *nothing is known* about their direct moral standing. This gives us some space for thinking about their indirect moral standing, and leads to 4 arguments that constitute 4 conditions for indirect moral standing. These are not necessary conditions but *sufficient* conditions for moral standing; there may be other sufficient conditions—and indeed other arguments for indirect moral standing.

First I will formulate my 4 arguments for indirect moral standing (as applied to robots) more extensively, unpacking the intuitions on which they build and their sources of inspiration, making explicit the morally relevant situations to which they relate, and elaborating them. I will also explore the relation to direct moral standing, explain how this indirect moral standing strategy applies the relational approach and responds to skeptic objections regarding the robot's properties, and briefly discuss the charge of relativism, which involves a reference to Deweyan pragmatism. Then I will discuss *how much* moral standing these arguments may establish and explore what these arguments mean for thinking about the moral standing of animals and—perhaps surprisingly and provocative—the moral standing of humans. These discussions involve a critical use of Levinas and an appeal to pragmatist ethics, in particular Deweyan ethics, which will enable me to draw on the concepts of otherness and moral imagination. With regard to contemporary work in robot ethics, the result constitutes a further and original elaboration of the normative implications of the relational approach proposed by Coeckelbergh and Gunkel, and a discussion of what this approach implies for the moral standing of animals and humans.

2 Four Arguments for Indirect Moral Standing of Personal Social Robots (4 Sufficient Conditions)

The first argument builds on the intuition that doing something "wrong" to a robot is not really wrong because of the robot but because of the human(s) involved. It is inspired by virtue ethics, which is becoming more popular in contemporary robot ethics (e.g. Sparrow 2020) and in ethics of technology more generally (e.g. Vallor 2016), and by the appeal to Kant in the literature (see Darling above). More specifically, the argument is based on Kant's argument for the indirect moral standing of dogs, which is related to his deontological ethics and his virtue ethics.

Let me explain this. First, Kant's argument for *direct* duties is clearly deontological, since based on the view of human beings as rational autonomous agents; this is why we have direct duties towards them. Kant's ethics, for example as expressed in the *Groundwork* (2005), is centered on the humanity of persons, which is in turn interpreted in terms of having rational capacities. And since (at least according to Kant) this does not apply to dogs and other non-human entities, he argues that we can only have *indirect* duties to dogs. We have indirect duties to non-human entities such as dogs not because we have a duty to *them*, but because, as Kant argues in *The Metaphysics of Morals*, we have a direct duty to ourselves: 'violent and cruel treatment of animals' is 'intimately opposed to human being's

duty to himself' since it 'dulls his shared feeling of their suffering and so weakens and gradually uproots a natural disposition that is very serviceable to morality in one's relations with other people' (Kant 1999, p. 564). What counts for Kant is our duties to other people and to ourselves, our dispositions and sensibilities, and our human capacity for (what we would now call) empathy, which helps in our relations with other people. This is not about the animals themselves; it is about human beings and their duties and sensibilities.

Second, however, as the reference to virtue in the citation about shooting dogs suggests, the full argument for indirect duties towards non-human entities is not only deontological but also connects to Kant's virtue ethics, which we can also find in *The Metaphysics of Morals*, where he discusses the relations between duties and virtue. Putting more emphasis on virtue, we can formulate the argument about dogs as follows: if we become habituated to behave cruelly towards dogs, that is, if we cultivate this vice and if our character becomes vicious, then we will likely become vicious towards human beings, which (and here we have deontology again) conflicts with our moral duty to respect humans as rational agents. Kant's deontological ethics is thus intertwined with his virtue ethics (how exactly is a topic that is beyond the scope of this paper; for a discussion of Kant's virtue ethics see for example Louden 1986). Note that from a (non-Kantian) consequentialist view, one could instead frame the problem as one that concerns consequences instead of duties or virtue: if one behaves in a cruel way towards dogs, one will behave in a cruel way towards humans. The challenge for someone wanting to protect non-humans, however, one needs to rely on empirical evidence to support such claims. This is difficult when it comes to effects of digital technologies. Deontological and virtue ethics arguments therefore carefully try to avoid going in this direction, for example when applying a virtue ethics argument to how to relate to robots (e.g. Sparrow 2020).

However, what concerns us here is not primarily what kind of normative ethics applies, but that Kant's argument about dogs is an argument for indirect moral standing, which renders it applicable to robots. If it is right, as Kant says, that we should not shoot a dog because we are a bad person if we do so and our behavior may spill over to humans—regardless of the direct moral standing of the dog—then whatever other reasons there may be to give moral standing to robots, we should not shoot a robot for the same reasons. *If* I accept Kant's argument for dogs, then I also have to accept the same indirect argument for robots; otherwise I am not consistent. Based on the brief consideration of Kant's view, we could formulate the argument as not necessarily being about the actual consequences to other humans, but as being about the virtuous character of the person. However, I propose to formulate this in a relational way: what is virtuous depends not only on my own assessment but also on that of others. Robots can then be given moral standing on the basis of this concern about the virtue of the human who might potentially do something that damages this virtue. For example, I may decide not to torture a robot because I care about potential corruption of my character, as seen by myself and as seen by others. (A similar argument can be made on the basis of not fulfilling one's duty and, as suggested, an alternative, non-Kantian argument could be made by framing this in terms of consequences.)

The second argument responds to the phenomenon that some humans in some situations get attached to robots and have feelings about or even for robots. It is not based on the capacity for moral agency of the human (human moral reasoning, decision, and actions) but on her capacity for moral patiency (what is due to humans). We should give moral standing to a robot if there is a human who has a (one-directional) relationship to the robot and cares about the robot (has developed feelings for the robot). The intuition here is again that what matters in such cases is the human, not the robot, but now it is the human in her capacity to care and feel, and another human who respects those feelings of care—thus respects the other *human* as moral patient. For example, I may decide not to hit a care robot that is comforting for a patient, not because I think that the robot has any moral standing at all based on its properties, but because I respect the feelings of the patient. And if I am the patient, I can ask others not to "mistreat" the robot based on my feelings. The analogy with Kant's dog is: I care about the dog; therefore you should not shoot it: not because it has intrinsic value and direct moral standing (maybe it has, maybe not, we bracket this for the sake of this argument), but because *I* care and I have (direct) moral standing as a moral patient and human subject.

The third argument is about play and collaboration: starting from the observation that humans play and collaborate with robots and value this and benefit from it, it says that we should give moral standing to a robot if, from the point of view of the human, the robot is a partner in play or necessary in a collaborative project, i.e. in joint action. This argument can be framed as a consequentialist one: what matters is the beneficial consequences for (other) humans. Useful analogies are a child playing with a dog and a dog for a blind person. The idea is that whatever other reasons there may be for not harming the dog, at least *one* reason and *sufficient* reason is that the child enjoys playing with the dog and the blind person needs the dog. Assuming that the joint play or activity is good, this makes it wrong to harm the dog, since doing so would cause harmful consequences: not for the robot, but for the humans. Again the question is if there would be harm on the part of the *human* involved. Another example: some people may desire to play games with robots or they may need the robot, e.g. as collaborators at work or as elderly persons. If this is the case, and if this collaborative activity is ethically good, then it would not be right to harm the robot in any way because of its extrinsic value and indirect moral standing as a partner in play or collaboration.

The fourth argument is a precautionary argument, which also starts from the human, now understood not so much as an agent or patient, but as moral doubting subject. Here the type of morally relevant situation which I try to do justice to with my argument is one in which it is not clear what the moral standing of the robot is. The reasoning is that if it is the case that there is any doubt, on the part of the human, about the robot's moral standing, then it is cautious and wise to give it at least *some* moral standing, since the moral cost and risk of being wrong is too high. Again this is an argument for indirect moral standing, since it is rooted in the doubt on the part of the human about the entity's standing, not directly in the properties and standing of the entity itself. Consider again our Kantian dog. Imagine the Kantian dog owner generally subscribes to the belief, common in his society and his time, that dogs do not have direct moral standing. But when living with the dog, he

may sometimes doubt about his dog's status as a mere thing. If this is the case, it is then better to err on the side of caution and give the dog some indirect moral standing, sufficient for preventing it from being shot. For us today, the case of the dog is now one to which we gladly apply a direct argument. But for many other animals, we are still not so sure. For these animals and in the absence of direct arguments, indirect arguments may well make the difference between life and death. And to come back to technology: an interesting case here could be that of interaction with an agent via the internet, without that we know if the agent is a human or artificial. If we really doubt about the moral and ontological standing of the entity we are interacting with, this argument would say that we better err on the side of caution and for example not hit a button that destroys or harms the agent in any way. (An interesting empirical experiment would be to test how people would actually act in such cases, at least provided that we can have technology that is able to create this illusion—which in turn invokes the discussion about Turing tests; however, here I focus on examining the argument.)

The rationale behind this argument, based on the risk that arises from not giving morals standing, is that humans have made serious mistakes in the past about the moral standing of other humans (e.g. people with a different skin colour) and animals (e.g. pets). Consider the analogy again: Kant, who, as we have seen, starts from a deontological view that restricts direct duties to humans, would not have considered giving direct moral standing to the dog he writes about, since he believed that they lack rationality and humanity. But today we think differently about this; many people today believe that dogs have direct moral standing. And most positions in animal ethics are critical of this argument since it fails to capture the intuition that some wrong is being done to the non-rational and non-human (Gruen 2017); they offer arguments for direct moral standing based on characteristics of the animals themselves, for example sentience. Now if it is the case that our moral intuitions are subject to historical change, would it not have been much better if the dog owners in Kant's time would have applied more caution when they doubted—even a little bit—the received view of the moral standing of their dogs? (And if we assume that their moral experience was not *radically* different than ours, they probably *did* apply some caution in practice and did not all and not always treat their dogs as mere things.)

Note that with regard to technologies like robots and voice assistants, this argument can be formulated as a requirement that moral standing be ascribed if the user has these doubts, but the condition can also be made stronger and more relational: for indirect moral standing to be ascribed, one could demand that moral standing is only ascribed if there is not only doubt on the part of one individual user, but also that in addition at some level of social organization there is intersubjective agreement between people (users, perhaps also including non-users) that there is doubt about the robot's moral standing. But how much agreement is necessary and at what level? Here is a proposal to deal with this, which applies the proposed precautionary reasoning: one could say that if one person (the one who interacts with the robot) has doubts, then this is sufficient for giving some moral standing to the robot, but that this assessment of the robot's standing can always be questioned by others (and perhaps: must be subjected to questioning by others); *if* there is widespread

intersubjective agreement that the robot has moral standing, however, then from a relational point of view this would constitute stronger reasons for giving moral standing, and hence would justify giving a more robust moral standing to the robot. (Note, however, that even this can be questioned again, a probably should be always open to questioning.) But other proposals could be made and one could discuss how much weight should be put on intersubjective agreement; this requires further work and needs to be anchored in theories concerning this topic. However, the main argument does not depend on how this issue is dealt with.

3 Further Discussion

None of these arguments rely on the claim that the robot in question has intrinsic value and direct moral standing. The arguments remain agnostic about that. In all cases its standing is derived entirely from valuing the relation we humans have to the robot (or a particular human has to the robot) and the related moral experience (e.g. perception of the robot, feelings about the robot) on the part of the human subject and subjects (plural) in a particular setting. This means that this kind of ethics is situation- and case-sensitive. The criteria are general but their application requires us to look into the particular situation. Whether or not indirect moral status is granted, all depends on the moral experience and actions of the humans, and on how humans perceive robots and how these humans are perceived (by other humans). This approach accords with traditions in ethics such as situational ethics, which takes into account the particular context instead of judging according to absolute moral standards, with moral theories that take seriously the role of emotions in the moral life (e.g. theories based on Hume), and in particular with Deweyan pragmatism, which rejects an ethics based on transcendent objective truths and instead sees moral theory as a method to address human and social problems based on human, collective experience (Fesmire 2003; Legg 2020).

One could object that this is not a strong basis for morality, or that it is wrong altogether to base morality on what people feel or agree (if this is a correct interpretation of this view at all), and that therefore it is better to rely on the firm basis of direct moral standing. In other words, there is the worry that this approach is relativist. I sympathize with this concern: surely morality should not be *reducible* to subjective feeling or social and cultural agreements. Now what relativism means (what varieties of relativism there are) and how it can be dealt with is a big philosophical issue by itself, to which I cannot do justice within the space of this paper. However, let me offer two responses. First, one could deny that this approach is relativist, or at least that it is one version of relativism. If relativism means the infamous "anything goes", then the pragmatist response is that this is not the case: even in the absence of absolute moral standards, moral experience teaches us that some ways are not good for people and their societies. Moral experience functions as a constraint and shapes what we feel and agree upon with regard to non-humans (and humans). This moral experience captures some wisdom and can guide us. Neither morality nor societies are static; there is moral learning. In the case of animals, this could mean pointing out that we learned to respect (at least some) animals more than we did in the past,

and that this now regulates our behavior towards animals—or at least some kinds of animals (e.g. pets). Therefore, it does not follow that anything goes, and a (pre)cautious approach towards other non-human entities is recommended, if only because if we fail to take such an approach, we risk to close off opportunities for collective moral learning. Second, one could accept that this approach is relativist in at least the (descriptive) sense that it claims that social norms are variable, but submit the consideration that historical and perceptual variation also applies to so-called absolute norms. One could admit that there is a degree of subjectivity and conventionalism involved in the actual, historically changed ascription of moral standing, and perhaps even sympathize with the yearning for absolute standards and making normative claims about moral consideration based on intrinsic properties, but point out that when people claim to rely on ahistorical, eternal and absolute moral standards and perceive certain properties in other entities, in practice (at the descriptive level) direct moral standing and even the properties turn out to be *also* a matter of perception and experience and agreement (as the relational approach suggests). For example, we now have very different feelings and agreed views about dogs than Kant and his contemporaries. While there is much discussion about the normative implications of these descriptive claims, it seems to me that at least one implication is the following: while moral experience has given us some insights, norms, and conceptual tools, we cannot fully trust our moral arguments about morality and general and non-humans in particular since (a) they turn out to be based on our moral intuitions and (b) since those intuitions vary historically and across cultures: previously we perceived animals very differently and agreed that they were mere things. Kant and other philosophers who denied moral standing to animals in the past also relied on a particular perception and agreement when they wrote about animals' lack of intrinsic properties and excluded them from the moral realm. Why would what we, contemporary philosophers say about the moral standing of entities be fundamentally different from them in this respect? We may believe that our views are non-dogmatic and based on reason, but so did Kant when he argued about dogs. Views have changed since then. Given this variation and change, then, and the untrustworthiness of our moral intuitions and of grand claims about absolute moral truths, I recommend precaution on the basis of a pragmatic and consequentialist argument: the reasoning presented in the precautionary argument may actually give a better guarantee, pragmatically speaking, that the relevant entities are treated well in the absence of any support for direct moral standing. (I will return to this point in the last section.) Moreover, nothing presented here prevents anyone from using reasoning that gives direct moral standing to animals or other non-human entities. *If* there are good arguments for giving animals or robots direct standing, the arguments presented give *additional*, indirect moral standing to these entities. In no way do they cancel out any arguments for direct standing. (Personally, I do not see how robots could be given any direct moral standing and believe the onus would be on the person who argues that they have; however, this is not the question in this paper. I remain agnostic about direct moral standing here for the sake of limiting the scope of my argument.)

Note also that it does not matter for these arguments whether or not what is happening on the part of the robot is "illusionary", e.g. whether the robot really

cares or has feelings. This question may well be crucially important for direct moral standing arguments. But what matters for these *indirect* arguments is only the reality of the actions of the human agent and the experience on the part of the human subject and subjects. In cases where there is doubt about the relevant intrinsic properties or even widespread agreement that such morally relevant properties are lacking (as is currently the case with robots), the proposed approach is helpful since it by-passes skeptic objections about the intrinsic properties of the robot (consider again the issue of consciousness) and circumvents the tricky topic of deception and related normative questions: whether deception is morally right, whether interaction with some robots counts as deception, and whether we should build robots that invite this phenomena. What the robot really "is" does not seem to matter for indirect moral standing, however important it may be for other reasons.

But is this right? Does this approach mean that intrinsic properties do not matter at all when it comes to indirect moral standing? Yes and no. Yes, they matter since in the background intrinsic (technical) properties may play a role in *creating the conditions* for the ascription of indirect moral status: the robot, say Teddy Bear 2.0. May fulfil one or more of these conditions in a stronger way than Teddy Bear 1.0. More than a doll, an autonomous and intelligent social robot might invite empathy, cruelty, feelings of attachment, it might become a partner for collaboration, and it might raise doubt through creating the illusion of being alive. No, because the indirect moral standing arguments are based on these human-related conditions, but do not rely on an account that explains how these conditions come into being. The arguments are not about the robot, but about how we perceive it and relate to it. In this sense, the answer is "no, intrinsic properties do not matter" for indirect moral standing, at least not in themselves. At best, they matter indirectly, since they influence the conditions under which indirect moral standing is ascribed. But according to the approach developed here, this is not important, at least not morally speaking. In the indirect standing arguments, the focus is on the appearance and experience, not on what produces the appearance and experience. The arguments are not based on, dependent on, the way a particular appearance and experience is produced.

The approach thus helps to overcome the epistemological problems indicated by Coeckelbergh and Gunkel (i.e. skepticism about how to know if entities really have particular properties), since the arguments remain ignorant about the intrinsic properties that produce the appearance. From a more naturalist perspective, the way intrinsic properties produce an appearance can be described. This is, after all, what social robotics and HRI are concerned with. The arguments presented here do not touch that. But the arguments also leave room for, and its approach is compatible with, the critical and more constructivist project of defining and analyzing the preconditions, the conditions that make possible our ascription of moral standing, which have to do with how the appearance is produced. Following Coeckelbergh, "what produces the appearance" need not be formulated in terms of intrinsic properties (Coeckelbergh 2012); the appearance is also produced in a relational way: through social relationships, through our use of language, and through our use of technology. The same is true for moral experience: a relational approach can inquire into its conditions of possibility, and for example study how language shapes our

moral experience of voice assistants. This is part of the project of a relational approach, understood as a critical philosophy.

Note that there may be other "relational" arguments for protecting robots, for example their cost and ownership (the obligation to respect the property of someone else) or the duty to avoid the environmental consequences of vandalism. These arguments are relational in the sense that they see the robot as part of socio-economic orders and ecological environments, and in the sense that they are also about humans relating to robots and the environment in particular ways. However, such arguments can be made for any objects, and are not specific to intelligent and autonomous robots who give the appearance of being alive. The specific indirect arguments offered in this paper are tailored to technological objects such as social robots that enable personal interaction and give rise to anthropomorphization, technologies that create specific phenomena and situations. But indirect arguments can also find more general application, and moral standing is a matter of degree. Thus, while the arguments can also made for teddy bears, they are particularly relevant to Teddy Bear 2.0. Since here it is likely that the conditions are fulfilled to a higher degree and thus that *more* indirect moral standing will be ascribed to it. This renders this account useful to contemporary and future robotics.

4 Indirect Arguments Applied to Animals

The arguments presented in the previous sections are developed for robots, but could be applied more widely to all kinds of entities, especially in cases when we are not sure about their direct moral standing. For example, they could also help the discussion about the moral standing of (non-human) animals. This could be done in at least the following way and in the following cases: in cases where there is doubt about the intrinsic properties of an animal (e.g., a particular fish or bird) and hence about its moral standing, we could at least give them indirect moral standing based on one or more of the arguments presented in this paper. This is a precautionary principle employed as a kind of meta-principle. I acknowledge that today we are sure that pets such as dogs and cats experience suffering. And most of us, at least in the West, give them a rather high moral standing, often based on our own moral experience. But not all animals fall in this category of moral certainty about their standing, and there are places where for example dogs are given less direct moral standing (e.g. where it is not uncommon to eat them). For instance, if we are not sure about what an octopus really feels and experiences (on the basis of the current state of the art in scientific research about octopuses), and hence cannot be conclusive about its direct moral standing, we can still give it some moral standing if one the following conditions are met:

1. We feel that if someone were to harm the octopus or were to be cruel towards it, e.g. torture it, we would consider the person a bad, vicious person and not trust the person to sit in the same room with our child, i.e. we would fear that the bad character or behavior carries over to humans. (The precise formulation depends on the moral theory used, as indicated previously.)

2. We have feelings of attachment and empathy towards the octopus and/or feel that we have a pet-like relationship with it (or any other kind of relationship for that matter).
3. We play together with the octopus, for example we collaborate in a game, or we need the octopus to perform tasks together (e.g. in an experiment); if someone were to harm the octopus, this would have bad consequences for us, humans.
4. We have doubts whether or not the octopus can feel pain and pleasure and whether or not we should give it moral standing. (And if we accept the point about inter-subjectivity, the condition could be added: "and others also have these doubts and we agree on having doubts about its standing.")

Note that number 3 would be congruent to the argument made by Coeckelbergh that non-humans such as animals and artificial agents can be drawn into the sphere of moral consideration (in that article: justice) on the basis of them being part of a human/non-human 'cooperative scheme' (Coeckelbergh 2009). This was also a social-relational argument, which, in the light of this paper, can also be interpreted as an argument for indirect moral standing. Moreover, Coeckelbergh's and Gunkel's use of Levinas to give animals moral standing as others that face us and call us to respond (Coeckelbergh and Gunkel 2014) can be interpreted as anticipating and implying an indirect approach to the moral standing of animals, in so far as it builds on the ethical relationship rather than direct, intrinsic properties of the animals. I will say more about this in the next section.

But *how much* moral standing would a robot, an octopus, or any entity for that matter, deserve on the basis of these criteria? For determining the degree of direct moral standing, we would need to find out to what extent the entity has the relevant moral properties. But how do we deal with this problem in the case of indirect moral standing? Here are some considerations. First, this would depend on the degree to which the conditions are fulfilled. If we do not relate much to the robot and if we have little doubts about its moral standing in the sense that we are pretty sure (but not totally sure) it has none, the moral standing assigned on the basis of these arguments would be low. Second, under these conditions its moral standing would probably be less than in the case of direct moral standing, assuming that direct moral standing is always higher than indirect moral standing. This assumption seems to hold, because (so a proponent of direct moral standing may argue) direct moral standing is all about the moral worth of the entity itself, something which is not addressed by arguments from indirect moral standing.

However, keeping in mind the lessons from the relational approach, our moral experience and more evaluation should not be reduced to a calculus, which takes distance from the entity in question. Asking the "how much" question itself is thus already a way to take distance from the entity. Moreover, there still seems to be something wrong about the very procedure of ascribing moral standing, even if directed at the moral worth of the entity, since this is done from a position of superiority as humans who have the exclusive right to do this ascription (next to having access to what the entity really "is" and to who or what is morally considerable—see the skeptical arguments again in Coeckelbergh 2012). Therefore, both arguments about direct moral standing *and* the arguments from indirect moral standing I

presented here, should not be understood as an attempt to close off the critical project of questioning our questioning (about moral standing). It is not a "final solution" that deals with the question concerning indirect moral status once and for all. They are working criteria submitted for discussion and (to pick up another pragmatist concept) experimentation. For pragmatists, the question is not only and perhaps not even mainly if these arguments and concepts fail in theory; in ethics and elsewhere, the main question is and should be, according to that approach: do they work in practice? In this context this could mean: given that direct criteria often in practice do not offer sufficient protection to animals and humans (consider concepts such as animal rights and human rights, which are based on direct arguments), could indirect arguments work better? This is not a question that can be answered in theory. But the question regarding humans leads us to the problem addressed in the next section: what about the indirect moral standing of humans?

5 Indirect Arguments Applied to Humans

Finally, it seems that there is one way in which intrinsic properties and direct moral standing still play a role in the indirect arguments presented: not with regard to the properties and direct moral standing of the robots or the animals under consideration (we are agnostic about this here), but the properties of humans. It seems that throughout the arguments, I have assumed not the direct moral standing of robots (I remained agnostic about that) but the direct moral standing of the *humans*. The arguments are based on taking seriously and respecting human beings as feeling, relational, social, playing, thinking, and doubting beings. It seems that it is assumed in this paper that these are intrinsic properties of human beings, which offer support for giving direct moral standing to humans. As in Kant's argument about dogs, direct moral standing arguments (about humans) form the basis for the indirect arguments (about other entities).

According to conventional (meta-)morality, this is not a problem at all. The (high and highest) direct human standing of humans and its foundational role is simply accepted as a dogma. However, from the point of view of the more critical, destabilizing project of Coeckelbergh and Gunkel, and taking on board lessons from Deweyan pragmatism, we must also critically question the basis of *that* view of (direct) moral standing: this view is not only anthropocentric but is itself also historical, is produced and constructed in certain ways (e.g. by means of language), and is not written in a kind of moral heaven but—at least at a descriptive level—seems to depends on what people do and how they solve societal problems. Seen from the point of view of absolutist moral philosophy, the proposed approach therefore runs the risk of relativism (see above). But seen from the point of view of relational and pragmatist thinking, the risk is that this account of indirect moral standing, based on the direct moral standing of humans, is interpreted as *absolute* and as constituting yet another dogmatic stance, meant to be ahistorical.

Moreover, there has been a lot of criticism of anthropocentric ethics. In thinking about animal rights and environmental ethics, in particular, it has long been asked why humans have a very high moral standing and non-humans not. This questioning

must be applied more widely and must certainly also be applied to the arguments I offered. Why do we humans use our own direct moral standing as a basis for the indirect moral standing of non-humans, and not the other way around? Why do we question the moral standing of non-humans, but refuse to question our own? We seem to apply two standards: one for humans and one for non-humans. This is perhaps understandable, given the long history of this kind of (anthropocentric) thinking, but it is not consistent. If our arguments about nonhumans start from an assumed instability and uncertainty about their moral status, then we should give our arguments about humans the same treatment.

If we follow this route, this raises at least two issues:

First, if direct moral standing accounts of humans—also relational ones—are more unstable than we thought, the question arises how to protect non-humans. I believe this can be done by fully embracing the view that the view of the human assumed in the 4 arguments is not an absolute view, but a working view of what the human is and what direct moral standing humans have, based on personal and intersubjective experience. This is not enough to convince the moral fundamentalist. However, from a pragmatic point of view it is sufficient to support the protection of non-humans in practice, especially if intersubjective agreement on indirect moral standing can be achieved.

But we should also ask the question: given the instability identified, can we protect *humans* in another way than by using direct moral standing views? Can we apply *indirect* moral standing arguments to them, and if so, what does that mean? This is not only a philosophical problem. Imagine that in practice the protections based on a direct account fail because, for example, a person or a group of persons really started to believe—perhaps under influence of propaganda—that a particular group of people are not humans or have a lower direct moral standing. Is there nothing that can be done, in terms of philosophical arguments, to justify treating them well, if we cannot appeal to an account of direct moral standing under such unfortunate circumstances? Imagine that such a person (influenced by propaganda) meets an other and, in spite of what her ideology tells her about the direct moral standing of that other, in her meeting with an other feels that something is due to that other and starts doubting whether that other really is a non-human being or a human being of lower moral standing, as her ideology prescribes. If there is no agreement on direct moral standing of that other, is there nothing she or we (from a third person perspective) can do to philosophically justify protections of that other motivated by this moral experience? Is the feeling this person has *just* an intuition or (from a third person point of view) "anthropological" observation, which is not relevant to morality, or does it have *moral* worth and can it be supported by moral reasoning?

An indirect approach is able to take this case seriously and deal with it. If for whatever reason the direct moral standing view fails in practice (whatever its worth may be from an eternal, god-like point of view), then it seems a good idea or even necessary, pragmatically but also morally speaking (in the sense of considering the moral consequences and risk of not doing so), to support the dignity of human beings by means of arguments for indirect moral standing. Furthermore, indirect arguments can be used in addition to direct ones. Nothing said here excludes or opposes reasoning about direct moral standing. Applying the arguments and criteria

developed in this paper to humans means that *whatever other reasons there may be* to protect human beings and not harm them, we should definitely ascribe moral standing to the "other" under consideration under the following conditions:

1. If not doing so would lead to a bad character (vicious character) or bad behavior towards other humans (or, with Kant: to not doing one's duty). Formulated in terms of cruelty: whatever other, direct moral standing reasons there may be for avoiding cruelty, being cruel to *one* human being may harm *your* character and lead to cruelty towards *other* human beings. This moral risk is not worth taking. If this argument is good enough for protecting the Kantian dog, it is good enough for us, and useful in case the direct argument does not work and is not implemented.
2. If we feel that we have a relationship with that other human being, feel empathy towards that person, and so on. Again the point is that this is already *sufficient* to justify moral protection, whatever other (direct) arguments there may be for protecting them and not harming them.
3. If we enjoy play or collaboration with that other human being and need that human being in and for our joint action.
4. If we doubt whether or not that other human being has moral standing, we should give the human being the benefit of the (moral) doubt.

The latter condition sounds awkward since usually we are certain about this, but might actually have very practical applications: consider again the case of the person influenced by propaganda (say in an authoritarian regime in the context of war) who has learned that the other is not a human being or has a very low moral standing, but starts doubting when confronted with a concrete human being that was supposed to have no direct moral standing. Hopefully this then leads to a change of opinion and the human being is seen as having direct moral standing. But if this does *not* happen, then reasoning that gives indirect moral standing may at least protect that human being.

And if this example seems remote (because we are privileged and lucky not to live under such circumstances), consider the technological sphere again, which was the starting point of the present philosophical investigation: with current and near future technologies such as robotics and AI, it is increasingly possible that we "meet" an entity of which are not sure if it is a human being or not. If we have no further way to test whether it is an entity that has moral standing (human or other), this presents a huge problem for direct moral standing criteria, which are based on the binary human being/non-human being and on intrinsic properties of an entity. In such cases, there is too much uncertainty, so the argument for direct moral standing cannot take off. But the uncertainty and instability in cases of, say very human-like humanoid robots or artificial agents that appear human, is not a problem for the indirect criteria offered here, since we can apply an indirect moral standing argument based on the precautionary principle: whatever the intrinsic properties of the entity (again, how can we know and how can we be sure? This is precisely a case where there is *doubt*), if we have any doubts about its moral standing, we better be cautious and treat it well.

Of course, intuitively one may assert that we should treat human beings as having a higher moral standing than Kantian dogs. Personally, and in the sphere of opinion and beliefs, I endorse the direct moral standing view that there is a moral obligation that we should always treat human beings well, based on their being-human, regardless of the moral standing of other beings. But we are not concerned here with opinions or moral beliefs as such but (1) with the *justification* of our moral beliefs and (2) with philosophically exploring what it means—against Kant—to take seriously the pragmatic and consequential aspects of moral beliefs and moral reasoning and explore the normative implications of a relational approach. Therefore, if it turns out that the direct moral standing reasoning that underpins this dogma is more philosophically slippery than many people assume and if there are practical circumstances under which direct moral standing does not work because it is not accepted or not implemented (for example when an authoritarian regime manages to convince many citizens that some people have less direct moral standing) or when there is too much uncertainty about a particular case, then it is safe—pragmatically *and* morally speaking—to use at least *also* indirect reasons.

These indirect arguments are not categorical but hypothetical. And they are not only about rationality but also have to do with feeling and relations. This is why in the Kantian tradition they are not considered as belonging to human morality. According to Kant, 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, p. 127). He thought we do not have direct duties towards them, only indirect duties. But here I moved beyond Kantian morality, taking seriously the pragmatic challenges of ascribing moral standing in a concrete relational and situational context. However, I do not claim that direct moral standing makes no sense or that rationality should play no role in morality. My point is that whatever direct arguments there may be for the moral standing of humans, indirect arguments can already do a lot of work to justify treating human beings well—especially in case direct moral standing arguments break down, do not convince, or are not implemented. And since they may help to protect humans (and non-humans), they are also moral, though not in a Kantian sense but in their consequences for the entities in question.

We may hope, of course, that usually justification is not needed, and that all this remains a "philosophical" exercise. We may hope that people treat others well, without any need for philosophers and others to debate about moral standing. We may also hope that in the future it will always be clear if an entity that confronts us is a human or not. But given our historical experience and the new possibilities of technologies such as robotics and artificial intelligence, the imaginary examples sounded too familiar and real; let us not hope too much. We need to create adequate conceptual equipment to deal with the challenges ahead.

This could mean that we need to rely on a wider variety of theoretical resources than usually employed in normative ethics. For example, keeping in mind Gunkel's contributions to the discussion about moral standing of machines (e.g. Gunkel 2012), which aim (among other things) to break down binary thinking with regard to humans and machines and use Levinas along the way, not only the relational approach as such but also arguments for indirect moral standing of humans seem to

accord with the thinking of Levinas. Interpreted as a relational philosopher by Gunkel and Coeckelbergh, Levinas (1969) argued that instead of establishing moral consideration on the basis of properties of individuals, we should put the ethical relationship before the ontological status of humans: the face of the other (or Other, as Levinas would say) interrupts and faces me in the encounter, asking for a response. Levinas developed this conception of ethics in the post-war period; he might have been thinking of a situation that can happen in a war (see also my example above): if we are confronted with another human being in an encounter, if we are *faced* by another human being, then the point of ethics is not to reason about its moral standing, but to respond. Ethics is first, not ontology. Here this means: the (moral experience of the) encounter and the relationship are first, not what that other (really) is.

However, using Levinas for thinking about indirect moral standing, more specifically for thinking about the indirect moral standing of humans and robots, is not without problems. For a start, Levinas's account cannot easily be applied to *robots*, since it was meant to apply to humans only. In response, Coeckelbergh and Gunkel (2014) have questioned the anthropocentrism of Levinas's approach and in the tracks of Derrida used Levinas against himself (Gunkel 2012, p. 182): while Levinas himself only considered human others, Coeckelbergh and Gunkel have used Levinas to defend a relational view with regard to animals. Instead of asking what properties animals have, Levinas's thinking invites us to ask the question how we can respond to the "face" of animals. Gunkel has even gone so far as to ask the question about machine others (Gunkel 2012). This can be seen as an extension of, and in any case a revision of, Levinas's own account, which was meant for human others and human ethics. Taken together, these Levinasian proposals radically question the very question of direct standing itself. But what exactly does that mean for humans, seen in the light of the previous discussion concerning direct and indirect standing? Does Levinas offer a direct or indirect argument for the moral standing of humans? Can his view of ethics be used to support an argument for *indirect* moral standing at all?

On the one hand and at first sight, Levinas's ethics seems to enable an argument for indirect moral standing. If the ethics of the encounter comes first, before any theory, then this could be interpreted in a relational way and as relying on indirect moral standing: moral standing is not about the essence of a human being and moral reasoning based on this, but about a situation and relationship in which the other appeals to my response-ability. I have to treat the other well not because theory tells me so, but because the other invites me to empathize in a concrete situation. This can be translated as constituting, or at least being compatible with, and argument giving the other "indirect" moral standing, in the sense that if take an ethical stance à la Levinas, I do not give moral standing on the basis of intrinsic properties of the other but on the basis of my relation to the other, or rather on the basis of my encounter with the other, who almost forces me to relate: who interrupts me and demands a response. The encounter is an experience which creates and appeals to my feeling that something is due to the other. I am responsible in the sense that I am called to respond. This, not the properties of the other, is the basis of the moral standing given to (and perhaps: demanded by) the other.

On the other hand, however, I suspect that in the end Levinas also relies on direct moral standing, in particular (and like all arguments for indirect moral standing

offered so far) on the direct moral standing of the human: the moral standing of the subject who is called upon, perhaps, and certainly the moral standing of the other. At the basis of the Levinasian view there seems to be an absolutist fundament, a view of the human being as Other with capital "O". In so far as there is less emphasis on the relationship and more on this absolute fundament in the Other, Levinasian ethics is extremely direct: rather than the relationship, it is the other's property of otherness alone (and perhaps also the humanity of the one who is called to respond) on which the ethics is based. This aspect of Levinasian thinking—at least if my interpretation makes sense at all—sounds like an argument based on direct moral standing. However, for the purpose of supporting the arguments in this paper, we do not need that dimension of Levinas's view for the arguments to work. If necessary, we can interpret Levinas in an "indirect" way and use his other-oriented ethics to try to move beyond an anthropocentric ethics—something which Gunkel and Coeckelbergh already suggest. It suffices that we take seriously the human moral experience of the situation and the encounter with the other (entity) as experienced by the human; this is where the 4 arguments start and end. And in order to take seriously this human experience, assuming some kind of direct moral standing of the human subject that ascribes moral standing seems unavoidable. But this does not seem to preclude that the ethical relation is directed towards the other. It seems that if we add Levinas to the account provided, we can both hold that (1) arguments for indirect moral standing are *based on*, *anchored in* the experience of the subject of moral consideration, rather than the intrinsic properties of the object of moral consideration, but (2) that this consideration is *directed towards* the other in a concrete situation. With Levinas, the relational approach becomes not other-based (because we do not consider the intrinsic properties of the entity) but other-directed. (For the purpose of developing this idea, further discussion about the relation between virtue ethics and Levinasian ethics would be helpful. However, this is beyond the scope of this paper.)

This emphasis on the concrete situation does not mean that in our indirect moral consideration we are necessarily limited to our immediate surroundings—a classic criticism voiced against moral particularism and relational approaches such as ethics of care, which could also be used against a Levinasian ethics, at least if facing others is understood as facing others that are near to me. But if we draw on the pragmatist tradition, this criticism can be answered by appealing to the notion of moral imagination, which has been proposed by Fesmire (2003) in the context of his interpretation and development of Deweyan ethics and which could be used as follows to expand the account offered so far: moral imagination can take us beyond the immediate situation, both in time and in space. All conditions in the arguments for indirect moral standing can be formulated in a way that takes into account the role of moral imagination in expanding our moral horizon from the present and local situation to possible future situations and situations elsewhere. We can say "If condition x is in place" but also "If we could imagine that". For example, we could formulate the condition about having a relationship with an entity as follows: "If we can imagine having a relationship with the entity, developing feelings for the entity, having empathy for the entity, etc., then we should give that entity indirect moral standing". This would widen the circle of moral concern from our immediate relations and situations to the future and

to potentially the entire planet and the universe. This is thus a way of answering the charge of particularism or bias towards those entities that are near us. If we plug in the concept of moral imagination, arguments for indirect moral standing and a relational approach do not necessarily imply that one is only concerned with the here and now, and with entities that are closely related to oneself. Using moral imagination, we can expand our feelings, our cooperation, our doubt, and so on. The conditions indicated could then also be met for entities that are further away from us. But note that even if their scope is expanded in this way, they are still indirect arguments, being not directly dependent on intrinsic properties of the entity as such and instead being all about the relations we have or *could possibly have* with other entities—including other humans—elsewhere and/or in the future.

Finally, does the view that humans have indirect moral standing undermine the arguments for indirect moral status of robots and animals? They seem to do exactly that, since for example the Kantian argument relies on direct moral standing of humans: it is the moral starting point for evaluating the moral standing of non-humans. This direct moral standing, in the form of the moral worth of human beings as rational agents, is seen by Kant as absolute: it is beyond questioning and not an empirical matter. When the direct moral standing of humans is questioned, then it seems we have no longer a basis to make that kind of argument. Here are at least two potential replies. One is to claim that we do not need absolute direct moral standing at all. It could be argued that "direct" moral standing is relational, and that we do not need anything more to get the arguments for indirect moral status going. All we need is a non-absolutist view of the human being as social, feeling, doubtful etc. but without making any absolute and foundationalist claims about this. Another response is to avoid the problem by saying that indirect moral standing comes in *addition* to direct moral standing. Nothing said here is conclusive about there being no direct moral standing at all. The arguments for indirect moral standing in the previous sections are agnostic about that, and some of the reflections in this section are critical and doubtful. But it is not necessary to claim here that there is no such thing as direct moral standing—of robots, animals, or humans—for the arguments for indirect standing to do their work.

The first option is worth further development. One could explore whether it is sufficient to rely on more contingent, relational criteria for giving humans moral standing in order to support the view that in some situations robots and animals have indirect moral standing. This may, in the end, collapse the very distinction between direct and indirect moral standing, since a truly (or more radically) relational view would presumably always consider the moral significance of both parties in the relation. My comments on Levinas went in that direction. More generally, one could further explore whether humans can live together on the basis of a less absolute, more relational understanding of themselves and their moral standing. This is a philosophical project on its own. The discussion in this part of my paper suggests that interpretations and applications of the philosophical tradition of pragmatism, the ethics of Levinas, or virtue ethics could help with this. But it is time to conclude.

6 Conclusion

The inquiry presented in this paper, which—in response to phenomena in personal social robotics—aimed to offer an account of moral standing that is not based on direct moral standing and that explores the normative implications of a relational approach to moral standing delivers at least two elements, which are important for ethics of robotics but also can be applied and discussed beyond that field:

First, it offers arguments for giving robots indirect moral standing under some conditions. This may be helpful to the discussion about moral standing of robots and the ethics of human–robot interaction, since it gives some concrete reasons for why robots may deserve moral consideration at all. These reasons, based on their indirect moral standing, are far less controversial as proposals for direct moral standing, which most participants in the debate do not agree with. My arguments respond to the intuitions that (1) people's experience of robots also should count "somehow" in moral thinking about robots, and (2) that there is some truth in the relational approach but that it is hard to figure out its normative implications. The paper makes a clear proposal concerning how exactly robots could receive moral standing and how the relational approach can support more practical and normative arguments for (indirect) moral standing. It thus constitutes a helpful intervention in a debate that tends to be focused only on direct moral standing and that is often inconclusive, confused, and unsure about the normative consequences of a relational approach. By proposing an indirect account of moral standing, it becomes clearer what a relational approach means in terms of *normative* guidance, as opposed to philosophical insights about how we think about moral standing and about machines. Moreover, the appeal to Deweyan pragmatism adds to the intellectual resources presented so far in the literature that develops the relational approach, and helps to respond to the charge of relativism and the problem that Levinas's account seems to limit moral consideration to the immediate situation.

Note that this paper was focused on a particular kind of robots, but its arguments could also be applied more widely to all kinds of technologies, including AI, provided that such technologies create conditions that are similar to the ones picked up by the arguments presented in this paper. For example, as I have suggested in one of my examples: if there were doubt about the moral standing of an artificially intelligent entity (in whatever form), then this would get the precautionary argument for indirect moral standing going. Since such conditions might arise in the future when higher degrees of narrow artificial intelligence are reached (e.g. when through improved machine learning and natural language processing it becomes more difficult to know if a text-mediated and web-mediated encounter involves a human being or not), arguments about indirect moral standing might come in handy.

Second, this paper also shows that this way of reasoning and this approach have also implications for thinking about animal ethics and, ultimately, reach deep down into how we think about humans and the fundaments of their moral

standing. That there are implications beyond ethics of robotics was already an emerging insight in existing work on the relational approach, especially with regard to thinking about animals. But this paper has explicitly touched upon the controversial question regarding the moral standing of humans and has used Dewey in addition to, and in response to, Levinas (who already figured in the discussions about the relational approach). More work is needed to develop this dimension of the paper, perhaps by further connecting robot ethics with (other) ongoing discussions in moral philosophy, epistemology, philosophical anthropology, and related fields. The focus here was on the moral standing of robots. But, as always, thinking about robots is not just about technology; it is also about us.

Funding Open Access funding provided by University of Vienna.

Data Availability Not applicable.

Code Availability Not applicable.

Compliance with Ethical Standards

Conflict of interest The author declares that they have no conflict of interest.

References

Bartneck, C., Croft, E., & Kulic, D. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1,* 71–81.

Breazeal, C. L. (2002). *Designing sociable robots*. Cambridge, MA: MIT Press.

Bryson, J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins.

Coeckelbergh, M. (2009). Distributive justice and co-operation in a world of human and non-humans: A contractarian argument for drawing non-humans into the sphere of justice. *Res Publica, 15,* 67–84.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moralconsideration. *Ethics and Information Technology, 12*(3), 209–221.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. New York: Palgrave Macmillan.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology, 27*(1), 61–77.

Coeckelbergh, M., & Gunkel, D. J. (2014). Facing animals: A relational, other-oriented approach to moral standing. *Journal of Agricultural and Environmental Ethics, 27,* 715–733.

Darling, K. (2012). Extending legal protection to social robots. *IEEE Spectrum*, 10 September 2012. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots. Accessed 22 June 2017.

Darling, K. (2017). Who's Johnny? Anthropomorphic framing in human–robot interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0*. Oxford: Oxford University Press.

Fesmire, S. (2003). *John Dewey and moral imagination: Pragmatism in ethics*. Bloomington and Indianapolis: Indiana University Press.

Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.

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

Gruen, L. (2017). The moral status of animals. Stanford Encyclopedia of Philosophy. Retrieved 30 September 2020 from https://plato.stanford.edu/entries/moral-animal/.

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

Gunkel, D. (2018a). The other question: Can and should robots have rights? *Ethics and Information Technology, 20,* 87–99.

Gunkel, D. (2018b). *Robot rights*. Cambridge, MA: MIT Press.

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

Kant, I. (1997). *Lectures on ethics*. In P. Heath, & J. B. Schneewind (Eds.), P. Heath (Trans). Cambridge: Cambridge University Press.

Kant, I. (1999). The metaphysics of morals. In *Practical philosophy*. Gregor MJ (Trans.) Cambridge: Cambridge University Press.

Kant, I. (2005). *The moral law: Groundwork of the metaphysic of morals*. H.J. Paton (Trans.) London and New York: Routledge.

Kant, I. (2012). *Lectures on anthropology*. In R. B. Louden, & A. W. Wood (Eds.). Cambridge: Cambridge University Press.

Legg, C. (2020). Pragmatism. Stanford Encyclopedia of Philosophy. Retrieved 30 September 2020 from https://plato.stanford.edu/entries/pragmatism/.

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

Louden, R. B. (1986). Kant's virtue ethics. *Philosophy, 61*(238), 473–489.

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

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics* (pp. 205–222). Cambridge, MA: MIT Press.

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

Sparrow, R. (2020). Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00631-2.

Sullins, J. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6,* 23–30.

Suzuki, Y., Galli, L., Ikeda, A. et al. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports, 5,* 15924. https://doi.org/10.1038/srep15924

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. New York: Oxford University Press.

Złotowski, J., Proudfoot, D., Yogeeswaran, K., et al. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics, 7,* 347–360.