Robots and the Limits of Morality

Raffaele Rodogno, filrr@cas.au.dk

To appear in: M. Nørskov (ed*) Social Robots: Boundaries, Potential, Challenges*. Ashgate

http://www.ashgate.com/default.aspx?page=637&title_id=1219171879&edition_id=1219254134&calcTitle=1

ABSTRACT

In this chapter, I ask whether we can coherently conceive of robots as moral agents and as moral patients. I answer both questions negatively but conditionally: for as long as robots lack certain features, they can be neither moral agents nor moral patients. These answers, of course, are not new. They have, yet, recently been the object of sustained critical attention (Coeckelbergh 2014; Gunkel 2014). The novelty of this contribution, then, resides in arriving at these precise answers by way of arguments that avoid these recent challenges. This is achieved by considering the psychological and biological bases of moral practices and arguing that the relevant differences in such bases are sufficient, for the time being, to exclude robots from adopting, both, an active and a passive moral role.

Introduction

Under what circumstances can robots properly be considered as moral agents bearing moral responsibility for their actions? Do robots have rights? Is it possible to harm or wrong a robot? I take it that most people (though, probably, robot-enthusiasts such as those reading these pages less so) will receive these questions with a certain amount of scepticism. Of course robots are *not* moral agents, and of course they *cannot* be harmed or wronged. They are after all just a bundle of circuitry and wiring typically wrapped in some poor human-looking form, if even that. There are no moral agents outside sane, adult human beings. And the only entities that can be wronged are those who have interests, such as, most clearly, sentient creatures like us. This is, of course, a crude answer. Yet it points us towards those issues that are crucial in order to answer the initial questions.

Electronic circuitry can be made to work wonders. Thanks to some such circuitry, we have built machines that beat humans at innumerable tasks. Why not allow, at least in principle, that we can build robots so sophisticated as to have whatever it takes to be part of our moral community, as agents and/or as patients? In this chapter, I shall not deny that this may in principle happen one day. I will, however, argue that that day is not today and that if that day ever comes, these robots would indeed be so similar to us that the decision to bring them about would be almost as ethically significant as the decision to bring human beings into the world.

Robots and Moral Agency

Under what conditions can anything, a child, a robot, or an adult human being, count as a moral agent? One answer to that question may go along these lines: anything that acts in accordance to what we would consider as the correct moral precepts would or should count as a moral agent. This, however, is not enough. We would not, and should not, consider as a moral agent something that just happened to act in the morally appropriate way out of sheer happenstance.

Consider next a context that offers minimal scope for moral action, such as one in which the only morally significant action is avoiding pushing a red button. Suppose you inculcate to your 2-year old child the notion that he never, under any circumstances, shall press the red button. Suppose that your child, whenever in the vicinity of the red button, diligently refrains (or avoids) pressing the red button. Once again, I hope to find my readership in agreement with the claim that we would not, and should not, consider this child a moral agent, at least not simply on the basis of his behaviour around red buttons.

What is it then that makes us (or whatever else) into a moral agent? In what follows, I will not afford a complete answer to this question. The partial answer that I am about to offer, however, is arrived at by interpreting a particular structural feature of the practice called morality. I am inviting us to consider morality as a specific realm of evaluation or judgement, opposed to non-moral realms of evaluation or judgement, such as the realm of aesthetics or etiquette. What is it, if anything, that characterizes moral judgements (or the substantive moral precepts that they express) that does not characterize other non-moral judgements?

The methodology used here to answer this question offers an interpretation of what it is that we do whenever we use key moral concepts in our judgements and evaluations, and in particular whenever we use the concept 'morally wrong'. The idea is that those who, for whatever reasons, cannot grasp this central part of our practice, i.e., the evaluation of something as wrong, cannot count as moral agents, despite their capacity to act in accordance to moral precepts.

The task at hand here does not amount to a semantic analysis of 'morally wrong', a term which on many accounts is considered to be primitive and hence not susceptible of further semantic analysis. It is rather the task of stating what conditions ought to obtain in order for us to declare that an individual competently uses the concept 'morally wrong'. On this particular view, concept possession is taken to be an ability that is peculiar to cognitive agents (Brandom 1994, Dummett 1993, Millikan 2000). Possessing the concept 'cat', for example, might amount to the ability to discriminate cats from non-cats and to draw certain inferences about cats. This understanding of concepts is associated to the conception of meaning as use. To learn the meaning of a concept is to learn how to use it, which is to make justifiable assertions involving it.[1]

Consider, for example, perceptual claims. On this account, at least part of the meaning of such claims is that we are justified in asserting them when we perceive things to be in a certain way and have no defeating evidence for that appearance. On this type of analysis, justified visual experience as of red, for example, is at least partly constitutive of the concept 'red'.[2]

[1] On this understanding of concepts, conceptual analyses are not understood as definitions stating *truth conditions* for the concept but rather as characterizations stating the *justification conditions* for that concept. In order to determine the meaning of a concept, we must start by identifying the set of conditions *C* that must be satisfied for a person to be said to have acquired a concept. The claim is, then, that for many concepts, the conditions *C* are uniquely determined by the justification conditions of the concept, which can then be said uniquely to determine the concept (Pollock, 1974, Ch.1).

[2] This, however, is not to deny the possibility that at the level of truth conditional definitions, 'red' is semantically primitive and 'experience of red' is defined in its terms or is analytic on the meaning of 'red'.

Going back to moral wrongness, on the view sketched here, the analysis of such concept would ultimately amount to the spelling out of the conditions under which a person would be justified in claiming that something is morally wrong. In the relevant circles, my concrete proposal is not new and is a version of what goes under the heading of neo-sentimentalism about morality. According to this view, the meaning of 'morally wrong' is (at least partly) constituted by our being justified in having certain emotions or sentiments.

Neo-sentimentalists disagree among themselves about what emotions are relevant to the analysis.[3] Hence, on one particular instantiation of this view (Gibbard 1992), to judge that Andy's hitting Ben is morally wrong is to say that it would be appropriate/fitting/rational for Ben and other impartial spectators to feel indignation at Andy and for Andy to feel guilt for her action. Others (Skorupski 2010), however, would reject that (justified) anger or indignation has a role to play in this context, while arguing that justified blame, either self-directed or other-directed, is what's ultimately constitutive of the meaning of 'morally wrong'.

Whatever the disagreements, in short, on this type of view, individuals correctly use the concept 'morally wrong' --make correct moral judgements-- when they master the normative attribution of certain emotions. Hence, for example, assuming that guilt partly constituted the meaning of 'morally wrong', it would follow on this view that someone who did not understand that guilt is appropriate only for actions for which one is responsible, will not thereby master the concept 'morally wrong'. An individual such as this (a child, perhaps) may well be inclined, for example, to judge that cats are evil because they torture mice. Similarly, on this view, when facing new or complex moral questions or situations, being experienced, imaginative, and skilful in attributing the relevant emotions will help one figure out whether an action is morally permissible or not.

Finally, we should note that it is *only* via the experience of certain emotions that we fully understand the meaning of 'morally wrong'. In order to show this much, let us return to the analysis of 'red' for a moment, and stress the fact that the meaning of this concept is partly constituted by justified *visual experience* as of red. Imagine a blind person who was, however, equipped with a high-tech device that communicated to her the light frequencies reflected by each of the objects she was touching with her index finger. Imagine also that this person had been taught that frequencies in the interval between 700 to 790 THz correspond to what people refer to as violet, frequencies in the interval between 600THz to 700 THz correspond to blue (and perhaps indigo), frequencies in the interval between 405 and 480 THz correspond to red, and so on and so forth. For as long as this person came into tactile contact with an object, she would be able to make most of the colour-related inferences that people with normal sight make. In many circumstances, then, she would appear to be a competent user of 'red'.

Yet, given the essentially visuo-perceptual nature of the concept 'red' in typical (seeing) subjects, the blind person will always miss the phenomenological aspect of 'red' and in fact will not be able to grasp the meaning of certain inferences that typical colour-concepts users make based on such phenomenology. Hence, for example, we may presume that it will be a mystery for this person why or how brighter colours (which she will identify simply as those having lower frequencies) are often thought as being happy colours though also as those colours we more easily get tired of. The blind person will not be able to understand these connections between colours, on the one hand, and mental states such as happiness, tiredness, fastidiousness, or being soothing, on the other, because these connections work precisely through the specific phenomenology of the different colours.

Now an analogous point applies to the neo-sentimentalist analysis of 'morally wrong'. Consider the very idea of being wronged or the idea that an injustice has occurred. These ideas are part and parcel of the phenomenology of emotions such as anger and indignation. When undergoing emotions such as these we are not simply undergoing the evaluation that certain norms have been violated. There are all sorts of norms such as aesthetic and prudential norms, and norms of etiquette, whose violation is not experienced in terms of

[3] Neo-sentimentalists disagree even more strongly with sentimentalists. While accepting that emotions play various roles in moral judgement, the latter deny that our moral judgements necessarily involve the normative attribution of emotions, as in "*x* is wrong only if it is *appropriate* or *rational* to feel emotion *e*."

wrongs or injustice. But when we experience anger and indignation (and perhaps guilt and blame), we experience the relevant norm violation, as it were, in moral colours, as precisely a wrong or injustice.

To return to our analogy with colours, just as the blind person, despite her high-tech device, will fail to grasp some important ways in which we use colour concepts, those lacking the relevant emotions will fail to grasp at least that part of the meaning of 'morally wrong' that is delivered to us through its phenomenology. For such individuals, it will be quite hard to distinguish violations of moral norms from other kinds of norm violations because the former will not be accompanied by their distinctive phenomenology, which include not only distinctive feelings but also specific sets of motivations (express blame, retort or punish, make amends). In other words, the emotionless will systematically miss a fundamental aspect of morality.

This point can be made more vivid through an example from literature. Think about what happens to Raskolnikov, the main character of Dostoyevsky's *Crime and Punishment* (1956), after killing the old lady and her sister in Part I of the book. The remaining six parts of the book are a vivid illustration of how it is (humanly) impossible for Raskolinkov to live with his crime. His guilt, his need to confess, and his need to atone make him at first physically ill and then almost insane. Raskolnikov's emotional life delivers and in fact constitutes the insight that he has committed a horrible deed. How can entities deprived of such emotions have access to *this* insight?

The claim being made here is that the concept 'morally wrong' and the moral practice that it animates *requires* an emotional base. This claim is crucial here, for, if it were true, robots will not be able to count as moral agents for as long as they could not be shown to feel the relevant emotions. The claim, however, can be read in two ways. On a stricter reading, each and every evaluation that a wrong or an injustice has occurred or would occur requires the occurrence of a relevant emotion (anger, indignation, or whatever emotion the neo-sentimentalist prefers). This, I think, is an implausible claim, which neo-sentimentalists do not and need not defend and which, therefore, I set aside.

On a looser reading, however, we could read the same claim as implying only that while a capacity to feel emotions is necessary to understanding moral concepts, an individual need not experience the relevant emotions on each and every occasion in which she engages in moral judgement. On this view, we would need to experience these emotions in our moral development, as we learn to grasp the peculiar nature of moral (as opposed to non-moral) judgements, and as we learn to ascribe moral wrongness correctly. Once we become acquainted with these tasks, however, while we will still *typically* experience such emotions when making moral judgements, we need not postulate their necessary occurrence. Individuals will have developed a sensitivity to these emotions, which will guide their judgements and their actions even in the absence of specific occurrences.[4] Hence, on the basis of your emotion-based experience, understanding and sensitivity you may, at least on occasion, morally condemn someone's action quite coldly, i.e., in the absence of any occurrence of anger, indignation, guilt or blame.

In short the fate of robots as moral agents hangs on their possessing the following:

> *Feeling.* The capacity to experience the relevant emotions and their actual experiencing of such emotions during the agent's moral development.[5]

[4] Our sense of (or sensitivity to) shame, for example, will often guide us towards certain actions and away from others. A shameless person is someone who lacks a sense of shame, and who thereby tends to act in ways that many people would find shameful (see Deonna et al. 2011 for a discussion of this topic).

[5] Coeckelbergh (2010) proposes a thesis that is on the surface similar but ultimately quite different from this. He argues in favour of the claim that robots be made to appear as if they experienced these emotions. His argument is based on the idea that we cannot be *certain*, not even in the case of human beings, that others experience emotions or, in fact, are conscious and have minds. For all we know with indubitable certainty, the moral games we play with other human beings are based as much on appearance as those we would play with robots that appeared to have the relevant emotions. In the next section, I will address this type of argument by showing that certainty is not and should not be

In fact, to be more precise, in accordance with neo-sentimentalism we should say that moral agency requires both *Feeling*, and, in addition to that, the capacity to make normative attributions of the emotions at issue. As illustrated above, it is not enough to show that an individual feels guilt. He or she (or it) must also be able to judge when it would be appropriate to feel it. Hence, in addition to *Feeling*, we must also consider the following as a requirement for moral agency:

> *Correct Attribution*. The capacity to make correct attributions of the relevant emotions.

Someone may at this point insist that moral agency requires only *Correct Attribution* and not, both, *Feeling* and *Correct Attribution*. After all, if someone were skilful at attributing the correct emotions, he or she (or it) would be skilful at making moral judgements. This line of argument should be attractive to those who want to defend the possibility that robots can become moral agents, for on the face of it, *Correct Attribution* does not seem to involve the capacity to feel emotions, which may indeed be very difficult to recreate in a machine.

There are at least two types of reply to this line of argument. First, if our capacity to make moral judgements has the phenomenological nature discussed above, then, emotionless individuals cannot quite grasp what it is involved in making moral judgements. We have defended that point by means of an analogy between moral concepts and emotionless individuals, on the one hand, and colour concepts and blind individuals, on the other. If this argument were not enough, we can help ourselves to another argument from the philosophy of mind and language, namely, John Searle's famous Chinese room argument (1980).

Suppose you are alone in a room following a computer program for responding to Chinese characters slipped under the door. You understand nothing of Chinese, and yet, by following the program for manipulating symbols and numerals just as a computer does, you produce appropriate strings of Chinese characters that fool those outside into thinking there is a Chinese speaker in the room. In a similar vein, I would take it that even if emotionless robots were able to reproduce successfully the moral judgements of a human community, via their capacity to make normative attribution of the relevant emotions (if that was indeed possible in the absence of *Feeling*), they would still fail to *understand* such judgements. They would, as it were, go through the motions without grasping their meaning.

Second, we should consider why, for what purposes, we would want to endow an entity that lacks the capacity to feel certain emotions with the capacity to make correct attributions of such emotions. In human beings, the latter capacity builds on the former. For example, while children may experience guilt before age 8, they acquire those capacities necessary to make correct attributions of guilt gradually, between the age of 8 and 12 (Harris 1989; and Tangney and Dearing (2002, 140-145)). If guilt were connected to morality in the way suggested by some neo-sentimentalists, we would come to make moral evaluations, among others, via our experiences of guilt. The correctness of such evaluations, however, depends on our capacity to learn to *feel* guilt appropriately. Suppose now that it was impossible to create robots that could feel guilt (or any other relevant emotion). What would be the point of equipping these robots with the capacity to make correct attributions of guilt? For one, these robots would not make moral evaluations via the experience of guilt.

It would make more sense to have these robots operate in the absence of any reference to emotion and operate rather with a list of proscriptions, itself devised on the basis of information concerning correct attributions of guilt in beings that *can* and *do feel* guilt. Robots operating in this way, however, will at most be accepted as agents acting within our moral community, if, that is, we trusted their capacities to recognize how to apply these precepts and how to solve the conflicts that may arise between them.[6] This is not to say,

what is at issue here. If we had good reasons to believe (rather than reasons to be certain) that an individual cannot experience the relevant emotions, but only systematically faked these emotions, we would not be able to relate to her as we would to a moral agent.

[6] Some caution is in order here. Many moral contexts can hardly be described as being structured by well-defined moral norms that stand in a clear hierarchical relation with each other. In such contexts agents are confronted with conflict of (moral) norms and the creative task of arriving at new norms in order to solve such conflicts. Here our capacity to feel

however, that they could coherently be considered to be moral agents. In the relevant respect, such robots would rather look like a version of the child in the red button example above with the difference that the robots' repertoire of proscriptions would extend beyond pressing red buttons. While such robots may be able to avoid performing morally wrong actions, we cannot infer from that that they grasp the meaning of this practice.[7]

Robots as non-derivative objects of moral consideration

Whatever one's stance about the status of robots as moral agents, the question of robots as appropriate objects of moral consideration remains open, for, after all, just as in the case of infants, to show that someone is not an active member of the moral community is not to show that he or she is not the proper object of moral consideration.[8] On the understanding of moral consideration at hand here, what is of interest is the possibility that robots be the object of *non-derivative* moral consideration.[9] In other words, showing that we ought to regulate our behaviour towards robots because of the rights or interests of those who own them or are otherwise attached to them will show that the robots are the objects of moral consideration only derivatively. The same is true of arguments showing that each agent should regulate behaviour towards robot because of the bad effects that failing to do so would have on the agent's character itself (Goldie 2010).

I will tackle the question of moral consideration by extending the neo-sentimentalist approach defended so far. Once again, my argument is deployed at the structural level. I will not, that is, discuss the merit of this or that particular moral norm. Rather, I will discuss the way in which some of the structural features of (this

the relevant emotions is essential in guiding us, which would in practice exclude those deprived of such emotions from taking part in such contexts, which, I assume, are not uncommon in our everyday life.

[7] This conclusion may of course be challenged by those who contest the claim that emotions play any role whatsoever in determining the meaning of morally wrong. This amounts perhaps to a version of the age-old disagreement in the history of ethics between so called sentimentalists, on the one hand, and rationalitsts, on the other. Kant's ethics is often considered to be one of the best examples of the latter tradition. It is generally difficult to solve such profound disagreements in a short space without begging the question. I will therefore refrain from attempting to do that here. It is interesting to note, however, that Kant's position may in the end not be so different from that defended above. Even according to Kant there is one moral emotion or feeling: respect (Achtung), which Kant sometimes seems to regard as the same as reverence (Ehrfurcht). Kant thinks that this emotion can motivate us. According to Sorensen (2002, p.110), however, the moral feeling plays a much more important role in Kant. Having argued elsewhere that that there is no experience without the capacity for pleasure and pain, in the third *Critique*, Kant argues that susceptibility or "the predisposition to the feeling for (practical) ideas, i.e. to moral feeling" is a condition for morality (Kant 1987, 5:265). There is no morality without the capacity for the specific feeling of respect or moral feeling. "As Kant will say later, it is a misunderstanding to think anyone could have a *duty* to acquire these sort of feelings, since they are "*subjective* conditions of receptiveness to the concept of duty" [(Kant 1996 [1797], 6:399)]." (Sorensen 2002, p.115) Here is, in other words, a Kantian version of the argument that emotionless robots cannot understand 'moral wrong'.

[8] Some authors (Gunkel 2014) seem to conflate the question of moral consideration with the question of (moral) rights. These, however, are slightly different questions. While being the bearer of moral rights involves being the object of moral consideration, the converse is not true: something may be the object of moral consideration without being the bearer of rights. An agent may, for example, have a moral duty to perform a certain action without a corresponding claim on behalf of any particular individual to have that duty performed. Hence, for example, charity may impose a duty to help others, without granting anyone in particular a claim or demand to be helped. (If such claim or demand existed, the action of those who addressed this claim or demand as a claim or demand owed to a particular individual would in fact cease to be considered as an instance of charity.) If this is correct, while it may be the case that robots are appropriate objects of moral consideration without being bearers of rights, the converse is not true. I therefore propose in what follows to stick to the broader question of moral consideration, so as to give robots their best chance to have moral status.

[9] By 'moral consideration' I henceforth mean 'non-derivative moral consideration', unless otherwise specified.

interpretation of) morality limit the content of all moral norms. To anticipate, I will show that our biology/psychology poses certain constraints on what can count as a moral norm. This has, in turn, repercussions on the question of moral consideration in general, and hence the status of robots as objects of moral consideration in particular.

On the neo-sentimentalist account there is no guarantee that societies will settle on an identical set of moral norms. While there are certain forms of behaviour that the majority of cultures understand and have understood as immoral (e.g., murder, stealing), clearly the proscription of many types of actions (and, perhaps, thoughts and feelings) is the expression of specific cultures or cultural forms (e.g., individualistic as opposed to collectivistic cultures). This should not come as a surprise for the sentimentalist, for emotional hermeneutics are clearly influenced by cultural processes.

What is of interest to us here, however, is not so much the degree to which moral rules differ across cultures, but rather the potential degree of variety in the content of moral rules. In other words, the question is whether moral rules can have any content whatsoever merely as a function of culture or whether there are limits to what may be taken to count as a moral norm. My proposal is that while cultural processes certainly play a large role in determining the content of moral norms, they also have limits, and in particular, limits that are imposed by the biological and emotional basis of moral practices.

Consider what explanation there may be for the fact that, while most cultures morally condemn taking the life of innocents, no culture has morally condemned (as intrinsically wrong) throwing pebbles in the sea or counting to ten. This is so for at least one reason: in order to exercise moral emotive responses, we must be able to see or sense a connection between what any moral norm proscribes and some disvalue that would accrue in the absence of such a norm. In other words, we must see the norms as connected to values in some way. Hence, while we can all understand that ceasing to live is typically a disvalue, the same is not evidently true for throwing pebbles in the sea or counting to ten.

Two further points must be noted in connection with this. First, given our emotive and cognitive nature, not only is the perceived existence of a value necessary to justify a norm, but the content of the norm must also be appropriately connected to the nature of the value. Hence, if ceasing to live is the disvalue, the relative norm must forbid taking lives, or promote the preservation of life rather than, say, proscribing the game of pushpin.

Secondly, as argued above, not all norm violations are considered as the appropriate object of moral emotions. This suggests that there may well be values that dictate norms other than moral ones. Hence, for example, while many agree that beauty is a value, and there are norms to the effect that beautiful things should be admired, failure to comply with such norms is not typically considered as the appropriate object of moral emotions. It wouldn't be appropriate for you to feel indignation at my failing to admire a beautiful thing nor would it be appropriate for me to feel guilt about it.

If this much is accepted, it follows that not all disvalue that accrues by violating norms regulating behaviour is *moral* disvalue. Not all values behind these norms are significant enough or of the right kind to count as moral values. The question, then, is what kind of dis/value can legitimately count as moral? I propose that *one* category of such values are those things that matter fundamentally or are of central importance to those who are the objects of moral consideration. Let us call these things interests. When it comes to human beings, there are a variety of views as to what kinds of interests we have. Some views are monistic, as, most notably, hedonism, according to which the only interest anyone can have is the experience of pleasure and the avoidance of pain. There are, however, pluralistic views, according to which, beside pleasure and pain, we have an interest in, for example, exercising our autonomy and freedom, maintaining deep personal relations, appreciating beauty, pursuing knowledge, etc. [10] The claim I am proposing, then, is that guilt, anger,

[10] Among pluralists one can further distinguish welfarists, who think that all these interests are parts or elements of human well-being, and non-welfarists, who, for example, would see well-being as a teleological value but would argue

indignation, or blame will tend to be considered appropriate when a violation of norms protecting what is considered an interest of this kind would occur.

In light of this framework, the question "Who or what is the object of moral consideration?" turns into "Who or what can have interests?" If, for example, the avoidance of pain is considered an interest, then, there is a pretty straightforward argument to the effect that moral consideration should be extended to all those creatures that are susceptible of pain and similarly with regard to other interests. At this point, then, the relevant question is whether robots can be understood as having interests in any sense. This question, however, raises an important epistemological issue. How do we know what interests, if any, robots may have?

In accordance with the frame presented here, we will be inclined to answer this question by considering what is known and intelligible to us, namely, human interests. In other words, just as in the case of animals and sentience, we will ask whether robots can have any of the interests that are intelligible to us from our human perspective, whatever our understanding of this list. In what follows I will not try to show, substantively, that robots (do not) deserve moral consideration insofar as they (do not) have recognizable human interests. The point I rather want to stress is the methodological/epistemological one. When considering the question of the moral status of an entity, human beings will not have much else to appeal to than what is intelligible to them in light of their biology in its cultural declinations, namely, whether such entities have interests and what kind of humanly recognizable interests these may be. In what follows, I will try to defend this particular anthropocentric epistemology by considering and dismissing other alternatives.

Environmental Ethics as an Alternative Framework

So far I have argued that having humanly recognizable interests grants moral consideration. I have not yet argued that interests are the *only* type of consideration that grants moral consideration. Perhaps there are other, non-interest based, types of consideration that do so. But what would these be?

In order to get some inspiration one could, for example, look at developments in environmental ethics. Research in this area is relevant here insofar as it aims at showing that non-human entities, such as the environment, have value and are the proper object of moral consideration. More in particular, there are positive efforts to understand the value of such entities independently of human interests (Rodogno 2010). The environment, that is, would be good even if it did not contribute to our well-being in any way, and, in fact even if no human being were there to value it.

Claims such as this, however, are not of the kind that we are looking for, as they may well be compatible with the idea that the environment is valuable insofar as *it* has interests of some kind. Now this kind of thinking is not an alternative to the one offered above. Rather, it confirms precisely the idea that our moral thinking is forced into adopting such categories. Perhaps an alternative way of thinking is offered by a holistic approach to the value of the environment such as Callicott's (1989, p. 25), according to which the *summum bonum* is the integrity, stability, and beauty of the biotic community. What ultimately matters on these views are the integrity, stability, and beauty of the whole, not the well-being or interests of individuals. How we are normatively to relate to the individuals (and their interests) composing that whole is regulated by the valuable features of the whole.

Such views raise an immediate question: Why is the stability and integrity of anything a value? Why not choose instability and disunity? And why think that these are *moral* values (the fact that Callicott includes beauty as part of the trio is not promising)? My guess is that if we tend to find at least some initial appeal in

that some of these interests are non-teleological and, therefore, not proper elements of well-being. For our purposes, however, we can set these disagreements aside.

the thought that stability and integrity are values, it may be because we believe that a certain degree of stability is necessary to the pursuit of our interests and then extend this framework to the environment. This, however, is no alternative type of thinking.

Even granting some initial appeal to the idea that states of systemic stability and integrity are valuable, another question remains. How do we, as moral agents, relate to such values? What norms follow from such values? Remember how we argued that moral norms must regulate behaviour in a way that is appropriate to the values they protect. What are we to do in order to respect the stability and integrity of ecosystems? The quick answer is of course that we have reason to *preserve* them. One way to read this claim is to take it as meaning that we should preserve them as they are now. But ecosystems are not unchanging systems. All of them were not in existence at some point in time, they mutate over time, and go out of existence at some point in time, even without any human interference. So, preservation of their *status quo* is not necessarily the right attitude.

Perhaps the claim should be read as meaning that we should see to it that we *not destroy* them. Rather than preservation, then, we should say non-destruction. The destruction or disappearance of ecosystems, however, is a fact of nature, independently of human activity. Intuitively, then, non-destruction cannot mean that we should actively do whatever it takes to extend the life of an ecosystem, just as valuing human well-being and life does not necessarily mean we have to extend human life *ad infinitum*.

Perhaps, then, the idea of *non-interference* is the best (deontic) interpretation of the claim that ecosystems are ultimate values. But the idea that we ought to refrain from interfering with ecosystems would be a strange one in many respects. First, by being on the same planet, we are bound to interfere with at least some ecosystems. We are part of this world and its nature as much as any other terrestrial being. Now, and this is a second reason against non-interference, all sorts of animals not only interfere but are constitutive elements of ecosystems. Their lack of an active role may cause the ecosystem to collapse. Third, human interference with an ecosystem is not necessarily detrimental. Some grasslands have been sustainably exploited for thousands of years (Mongolia, Africa, European peat and moorland communities).

Returning to our discussion of the moral status of robots, we should take this excursus into environmental ethics as an illustration of the kind of problems that are likely to arise if we tried to abandon the framework defended so far. First and foremost it seems indeed that our moral thinking is bound to the category of interests. If we try to unshackle it from this category, a number of difficult questions arise. If robots do not have interests, on the basis of what should they be the object of moral norms and consideration? Supposing that they did have non-interest based values, what would they be? And how could we determine the content of the norms that such values would dictate to us? Finally, would these norms be rightly understood as moral as opposed to non-moral norms?

A Final Hurdle

Before these difficult questions, one may conclude that the safer alternative is to accept the limits of our moral thinking and stick to the type of frame defended above. This, however, would be a hasty conclusion, for there may still be an alternative approach and, in the absence of that, we may want to reject the frame defended here on other grounds. In what follows, I will therefore present, in turn, some recent criticisms that have been put to an approach which, at first sight, seems to be very close to ours; and, then, present and reject a last alternative approach.

The Standard Approach is characterized as follows:

> entity x has property p
> any entity that has property p, has moral status s
> entity x has moral status s (Coeckelbergh 2014, p. 63)

On the face of it, our approach may be seen as an instance of the Standard Approach, for, on our approach, we would indeed claim that any entity that has the property of having interests, has moral status, and, even more strongly, that only those entities that have that property have moral status. We would then go on to examine whether robots display such properties and, for as long as we have reason to believe that they don't, we would deny that they are non-derivative objects of moral concern.

Both Gunkel (2014, pp. 120-121) and Coeckelbergh (2014, p. 63) believe that this approach falls prey to serious epistemological problems affecting the first and second steps. In particular both authors espouse a sort of scepticism that implies that we cannot be certain and cannot have indubitable knowledge of other minds, and in particular or whether other minds have the capacity to suffer or are otherwise conscious. This obviously affects the first step of the standard view, as many of the interests that we recognize are indeed dependent on sentience and/or on having a mind. The same kind of skepticism, however, affects also the second step. As Coeckelbergh (2014, p. 63) puts it:

> 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.

Our approach, however, can eschew these epistemological problems. Sure enough, we cannot, *with certainty*, make claims about sentience and other minds but we do not need to make such indubitable claims. On our approach it is sufficient to have justified beliefs about sentience and other minds. Inference to the best explanation would be enough here to justify our believing that other human beings and animals sufficiently similar to us have mental states. The argument would go as follows:

> Other human beings are very like me. They behave very much as I do in similar circumstances *and they are made of the same stuff*. When I burn myself it hurts and I cry out and wince. When other people are burned they do the same. I can thus infer that they are in pain too. There are multifarious such similarities. Put more generally, I know directly that I have beliefs, emotions, feelings, sensations and the like. So I am enabled to infer, on the basis of these multifarious similarities, that other people also have beliefs, emotions, experiences and the like. In short, I am entitled to infer that other human beings have as I do, an inner life and that it is very like mine. (Hyslop 2014, my emphasis)

Now, for as long as robots are not sufficiently similar to us in the relevant respects, i.e., made of the same stuff, we have no reason to believe (rather than indubitable knowledge) that they have minds and/or are sentient and, hence, cannot attribute to them the kinds of interests that give rise to moral status. Similarly, the claim that we, as human beings, have no real alternative other than conceptualizing moral status in terms of humanly recognizable interests is *not* an *a priori*, indubitable intuition read off a Book of Values, but an *a posteriori* critical interpretation of our moral thinking, open to challenge by other interpretations.

The troubles for the Standard Approach and its affiliates, may, however, not be over. On this type of view, there is an alleged unbridgeable *explanatory gap* between the attitudes of people that attribute mental states and engage affectively with robots, on the one hand, and the idea that robots, given their lack of minds, sentience, and interests, are "mere machines". When it comes to robots and their moral status there is, in other words, a gap between what certain people experience when engaging with them and our reasoning about them, between thinking and action, between belief and feeling. We may *think* about them as mere machines, when thinking about them in a scientific mode, and yet experience particular robots as "more than machines" and address some of them with "he", "she", or even "you". Coeckelbergh (2014, p. 64) goes on to write:

> 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?"

This explanatory gap, however, is not a necessary feature of the Standard Approach, not, at least, in the version defended here. As argued elsewhere (Rodogno Forthcoming b), there are a variety of ways in which we can conceptualize human-to-robot affective engagement that fall short of attributing irrationality or immaturity to the human. Hence, for example, we could hypothesize that, when engaging affectively with robot pets, individuals adopt a cognitive mode akin to that which is normally adopted in our engagement with fiction. Being emotionally engaged by robot pets would be akin to being emotionally engaged by a good novel or movie. Just as my sadness for Anna Karenina involves my *imagining*, *accepting*, *mentally representing* or *entertaining the thought*, *without believing*, that certain unfortunate events have occurred to her, my joy at the robot pet involves my imagining, accepting, mentally representing or entertaining the thought, without believing, that it is happy to see me. No need to misrepresent the world or be irrational about anything.

I suspect, however, that this type of answer may still fail to satisfy the critics of the Standard Approach. According to Coeckelbergh, philosophers adhering to this approach simply adopt the wrong kind of methodology, as they proceed by isolating robots from their contexts and situatedness, in the same manner in which scientists isolate the object of their study from its environment by creating artificial, experimental conditions. By proceeding in this way, philosophers lose sight of precisely the fact that the object of their study stands in specific relations with its environment. An elderly attached to a robotic pet, who treats the "pet" as a real pet or baby, for example, is already engaged in a relation with the robot that goes beyond the question of the moral status of the robot. At this point Coeckelbergh (2014, p. 70) suggests:

> Is not moral quality already implied in the very relation that has emerged here? … What needs to be discussed is that relation, rather than the moral standing of the robot.

On the alternative approach Coeckelbergh (2014, p. 65) proposes

> 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.

On this alternative approach there is no "correct" way of seeing the robot. Moral status ascriptions are at one level growing within the relation. This, I believe, is the point at which the Standard Approach and the approach defended here will stop accommodating the criticisms made by Coeckelbergh. Why think that the fact that an individual is attached to a robotic pet in the same way as he would be attached to a real pet or a baby makes the question of the moral standing of the robot irrelevant or superfluous? Is it because it is already so clear that for the individual who is attached to the robot the latter already has moral status? But is that correct? If the suggestion made above is descriptively accurate, it may well be that the individual in question engages with the robot in the same way in which we engage with fiction. Her emotions would be real but nonetheless resting on cognitive bases that are not regulated by a truth-norm. The fact that the robot pet actually has no interests and hence no moral status would not give her any reason to change her attitudes towards it or be less attached to it.

Even admitting that those attached to robot pets took the latter to have moral status, the question of the moral status of the robot would still be entirely pertinent. For one, in line with the view discussed above, we should understand what moral norms would follow from the fact that this particular individual is attached to this particular robot. Does it, for example, follow that you and I and everybody else now have reason to regulate our behavior towards the robot? The answer, I would imagine, is positive: we should regulate at least some of our behavior with regard to the robot. But in order to understand in what respects we should do so, we would need to understand whether the robot is a derivative or non-derivative object of moral consideration, for the type of norms that would follow in either case would indeed be different. Now in order to argue that the robot has non-derivative moral status, it would not be enough to show that its relation with one individual already involves "moral quality".

Morality is first and foremost a practical affair, a system of rules whose rationale is the regulation of behavior. It works through the establishment of norms that will at times restrict the agents' freedom to pursue their own interests. A stance according to which there is no "correct" way of seeing the robot and its moral status, which makes appeals to a supposed "moral quality", will fail to be sufficiently practical. As argued above, agents will fail to understand and, consequently, be moved by such appeals unless they could see how robots have recognizable interests. Yet they will, for the same reason, understand and be moved by claims to the effect that we should "respect" the robot for the sake of the person who is attached to it.

For as long as we have no reason to believe that robots have interests or other plausible alternative ways of thinking are presented, we shall rest content with the idea that we should regulate our behavior towards robots only derivatively, that is, for reasons having to do with the interests of those who are non-derivative objects of moral consideration.

Conclusion

In conclusion, then, our biology and psychology pose limits to human morality in two important respects: by excluding from the realm of non-derivative objects of moral attention anything incapable of humanly recognizable interests, and by excluding from the realm of moral agents anything incapable of feeling a certain range of emotions. As argued at the start, this is not to say that robots will always be excluded from our moral community in these two important respects. That question is not one I have examined here. The day in which robots fulfill the conditions for moral agency and moral patience outlined here, however, will be the day in which they will be emotive creatures capable of humanly recognizable interests. These robots would indeed be so similar to us that the decision to bring them about would be almost as ethically significant as the decision to bring human beings into the world.

REFERENCES

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Cambridge, MA: Harvard University Press.

Callicott, J. B. 1989. *In Defense of the Land Ethic: Essays in Environmental Philosophy*. Albany, NY:
SUNY Press.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12 (3). pp. 235-241

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

Deonna, J., Rodogno, R., Teroni, F. (2011). In Defense of Shame: The Faces of an Emotion. NY: Oxford University Press.

Dostoyevski, F. (1956 [1866]). *Crime and Punishment.* Constance Garnett trans. New York: Random House.

Dummett, M. (1993). *Seas of Language*. Oxford: Oxford University Press.

Gibbard, A. (1992). *Wise Choice, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Goldie, P. 2010. The Moral Risks of Risky Technologies. In Roeser, Sabine (ed.) *Emotions and Risky Technologies*. Springer. 127-138

Gunkel, D.J. 2014. 'A Vindication of the Rights of Machines'. *Philosophy & Technology*, 27 (1) 113-132

Harris, P.L. (1989). *Children and Emotion: The Development of Psychological Understanding*. Oxford: Blackwell.

Hyslop, A., "Other Minds", *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/spr2014/entries/other-minds/>.

Kant, I., (1987 [1790]). *Critique of Judgment*. Werner Pluhar, trans. Indianapolis: Hackett.

Kant, I. (1996 [1797]). *Metaphysics of Morals* in Kant, *Practical Philosophy*. Cambridge: Cambridge University Press.

Millikan, R. (2000). *On Clear and Confused Ideas*, Cambridge: Cambridge University Press.

Pollock, J. (1974). *Knowledge and Justification*. Princeton, NJ: Princeton University Press.

Rodogno, R. 2010. Sentientism, Well-Being, and Environmentalism. *Journal of Applied Philosophy* 27.1, 84-99

Rodogno, R. 2015. Social Robots, Fiction, and Sentimentality. *Ethics and Information Technology*. On-line First.

Searle, J., (1980). 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3: 417–57

Skorupski, J. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.

Sorensen, K. 2002. “Kant’s Taxonomy of the Emotions”. *Kantian Review* 6, 109-128

Tangney, J. P. & Dearing, R. L. (2002). *Shame and Guilt*. New York: The Guilford Press.