Machines and the Moral Community

Erica L. Neely

Abstract: A key distinction in ethics is between members and non-members of the moral community. Over time, our notion of this community has expanded as we have moved from a rationality criterion to a sentience criterion for membership. I argue that a sentience criterion is insufficient to accommodate all members of the moral community; the true underlying criterion can be understood in terms of whether a being has interests. This may be extended to conscious, self-aware machines, as well as to any autonomous intelligent machines. Such machines exhibit an ability to formulate desires for the course of their own existence; this gives them basic moral standing. While not all machines display autonomy, those which do must be treated as moral patients; to ignore their claims to moral recognition is to repeat past errors. I thus urge moral generosity with respect to the ethical claims of intelligent machines.

1. Introduction

A key distinction in ethics is between members and non-members of the moral community; this is the foundation for understanding how we should treat the entities we encounter in the world. Over time our notion of this community has expanded; those we take as non-members have changed, and the criteria used to make that distinction have also altered. Historically, as surveyed by Lorraine Code (1991), Charles Mills (1999) and Naomi Zack (2002), criteria such as intellect and rationality were used to separate white men from women and non-whites. Taken to be governed primarily by emotion rather than rationality, these people were seen as moral inferiors, deserving of lesser or no moral consideration.

Even upon conceding that rationality was not the exclusive preserve of white men, and so including women and non-whites as members of the moral community, many continue to deny moral standing to animals. Both contemporary thinkers (Scruton 2006) and earlier philosophers (Kant 1786/1996) see humans as having the moral high ground of rationality and consciousness. However, rationality criteria raise questions as to how rational a being must be to receive moral standing – there is a serious risk of excluding certain humans (such as infants) from the moral community which, as discussed by Peter Singer (2002), is unpalatable to many thinkers. Furthermore, our understanding of the biological similarities between humans and other animals makes it difficult to maintain a sharp distinction between them; various other animals seem to possess degrees of rationality and consciousness as well.[1] Such reasoning has caused many (Bentham 1823/1996; Taylor 1996; Singer 2002) to move to sentience as the criterion for moral standing: if something can feel pain, it is wrong to take intentional action to make it suffer unnecessarily.[2]

This is a large expansion to the moral community, yet of course many things continue to lack moral standing; an object such as a table or chair is not a member of the moral community, for instance, because it is not possible to cause moral harm to the object itself. Unless the object belongs to someone else, I can do what I wish to it; the only kind of moral harm that can be

[1] For instance, the National Institute of Health (2013) has recently designated chimpanzees as inappropriate for most forms of animal research, since they are our closest relatives and "are capable of exhibiting a wide range of emotions; expressing personality; and demonstrating individual needs, desires, and preferences." The sort of clear distinction between human and non-human animals once thought to exist is increasingly being challenged, giving rise to new ethical implications.

[2] Obviously there is clarification required to specify what constitutes unnecessary suffering and exactly how much moral standing animals have. However, sentience suffices to give them a foot in the door of the moral community, so to speak.

caused in this situation is harm to a person or persons who have a claim to that object.[3] As such, there is currently a strong ethical divide between living beings and non-living things. This has serious implications for the ethical issues pertaining to intelligent machines, since they too are inanimate objects. There is a strong temptation to classify them as similarly undeserving of any moral standing.

I will argue that the relevant criterion for membership in the moral community should not be understood as whether one can feel pain but rather whether something has interests. While sentience is certainly one way of having interests, it is not the only one. Using this criterion, I will show that certain kinds of machines are, in fact, members of the moral community; specifically, I will argue that they are moral patients.[4] Thus while we are correct in extending membership in the moral community to encompass humans and animals, we must further extend it further to include these machines.

2. Ethics and the Prevention of Harm

When conversing with people, one informal objection that frequently occurs to granting moral standing to a machine is the claim that you cannot "hurt" a machine. In essence, this is an internalization (and over-simplification) of the sentience criterion for moral standing. Ethics is often taken to involve the prevention of harm; as David Gunkel (2012) notes, the central question of moral patiency often is phrased as whether something can suffer. Hence if something cannot be harmed, many are reluctant to offer moral standing to the thing in question.

For humans, the harm generally involves some kind of pain. However, the ability to feel physical pain cannot be the only criterion for membership in the moral community. Consider a person with congenital analgesia, i.e., one who is unable to register physical pain. If someone were to step on his foot, he would not be able to feel any pain from the action. Yet it would surely be wrong if one stepped on him simply because one took a kind of perverse amusement in his inability to feel it. Stepping on his foot intentionally, without his permission, and without some kind of greater justification strikes us as wrong.[5]

This is not because the action caused pain (since, by design, it does not); sentience as construed to involve purely physical sensations is not sufficient to render this action wrong. Furthermore, even if we extend sentience to include mental or emotional pain, it is still insufficient; it is wrong to cause harm even if the victim is emotionally unmoved, such as when we see emotional dissociation in child soldiers or victims of abuse. The wrongness in our case stems from two key points. First, the action could cause damage, even if it does not cause pain. Second, since we have specified that the person does not give permission for the action, deliberately stepping on his foot violates his desire to remain unmolested; moreover, there is little justification for this violation.

---

[3] The ownership of an object could be the community as a whole, such as with public art installations. If someone were to destroy the Vietnam Veteran's Memorial, one could argue that it would cause harm to the public (which has a claim on the memorial) and is thus morally wrong. It would be odd to say that you had morally wronged the monument itself, however.

[4] I am concerned in this paper with what it takes for a machine to be deserving of rights and hence be a moral patient. I leave open the question of what it would take for a machine to have moral responsibilities and thus be a moral agent.

[5] This action might be justified if it were done out of a different motivation. Even if I lack his consent, deliberately stepping on his foot might be acceptable if it prevented a greater harm (such as stepping into the path of a vehicle.) However, this is a rather different case than interfering with another's body simply because it entertains me.

What is necessary for moral standing is not sentience per se but having interests; the person in our congenital analgesia example lacks sensation, but he retains interests. As it is possible to harm those interests, it is possible to harm him. To expand, consider John Basl's definition of interests as "those things the satisfaction of which contributes to [an individual's] welfare." (Basl 2012) This implies that a being can have interests without being aware of them. For instance, an ant may have an interest in not being stepped on and killed even if the ant is unaware of that interest; similarly for a person in a persistent vegetative state.[6] In each case, their welfare can be threatened regardless of whether they are aware of that threat. Of course, many times we are aware of our interests – we have ideas about how we wish to run our lives, and thus have some interest in those plans being followed; we may be harmed if our desires are simply ignored.

The notion of harm that is relevant for morality, therefore, moves beyond physical pain and hinges on the idea of disrespecting the integrity and autonomy of the individual. The possibility of the action causing damage, even if it does not cause pain, raises the idea of bodily integrity. At a minimum, beings have an interest in retaining sufficient bodily integrity for continued existence; anything which damages one's body threatens this interest. This interest can certainly be outweighed by other factors – I may consent to having my appendix removed because that particular violation of bodily integrity actually promotes my continuation under certain circumstances. Frequently in medicine we consent to actions which are extremely damaging to our bodies (such as chemotherapy) if the alternatives are worse.[7]

In addition to these dramatic cases, we consent to small violations of bodily integrity on a regular basis; it is clearly possible to overstate our commitment to it, since most people trim their fingernails or their hair or will pick open the occasional scab. Yet people are unlikely to see those actions as presenting any serious threat to continued existence. Hence one might argue that a minor harm, such as stepping on a person's foot, cannot truly be objected to on this basis alone. Indeed, I believe that the emphasis on bodily integrity dovetails with the desire to remain unmolested mentioned above; together they highlight the fact that we have certain wishes about the shapes of our lives.

By ignoring the person's desire not to be trod upon, the aggressor's action violates his autonomy. In much of ethics, autonomy is emphasized as an important good.[8] To cast it aside for no reason other than to satisfy one's own sadistic desires is to jeopardize the interest of the injured person in governing the course of his own life. Such an action may not cause physical pain, but it clearly causes harm to that person – it treats him as incapable or unworthy of directing his own actions, and views his desires as irrelevant and something that may simply be

[6] This is why it would, for instance, be wrong to take pornographic photos of a person in a persistent vegetative state; we believe that a person can be harmed even if he or she is unaware of it.

[7] One could also justify suicide this way for some cases, since my interest in bodily integrity could be outweighed by an interest in avoiding large amounts of suffering from a terminal disease, say. While we have an interest in bodily integrity, it is not the only interest that matters.

[8] We see this both in Kant (1786/1996) with the view of rational beings as ends-in-themselves and in Mill (1859/1993) with the emphasis on individual liberty.

ignored.[9] Although it is clear that sometimes a person's desires must, ethically, be overridden, we surely cannot ignore another's wishes completely. [10]

Hence while sentience certainly leads to having interests, it is not necessary for them: the joint properties of consciousness and self-awareness will also suffice.[11] Once a being is self-aware and conscious, it is aware of its self, can desire continuation of that self, and can formulate ideas about how to live its life.[12] It is possible to harm such a being by ignoring or thwarting those desires; one should not act against the being's wishes, therefore, without some overriding reason. The requirement of such a reason, however, is equivalent to granting the being at least minimal moral standing; one does not need to have a reason to destroy a chair, but one must provide such a reason to destroy a human. This holds true for intelligent machines just as much as for a person with congenital analgesia; they both have interests and desires, hence they both have basic moral standing.

One could object at this point that we have moved too quickly from consciousness to ascribing desires to a machine. Basl discusses the possibility of a machine which is conscious only insofar as it has the ability to experience colours. However, it has no emotional or cognitive responses to those experiences – it may experience blue, but it does not care. Basl claims that it would not be wronging such a machine if we were to shut it down. Similarly, suppose a being existed which could feel pain but had no aversive reaction towards it; furthermore, this is the only conscious experience the being has. In this case, the being would presumably have no interest in avoiding pain, and Basl believes that it would not be wrong to cause pain to it. In each of these cases, Basl argues, there is consciousness without moral patiency. Consciousness, understood as the ability to have sensory experiences, is not sufficient for having interests – instead one must have the capacity for attitudes towards those experiences. (Basl 2012)

Basl's view of consciousness is extremely limited, however. Steve Torrance (2012) notes that "To think of a creature as having conscious experience is to think of it as capable of experiencing things in either a positively or negatively valenced way – to think of it as having desires, needs, goals, and states of satisfaction and dissatisfaction or suffering." I believe that this robust notion of conscious experience is more in line with what we mean when we consider conscious machines. Basl is correct in arguing that the liminal cases he describes are likely not instances of moral patiency. Furthermore, there is importance to considering these instances, since it seems probable that the first conscious machines will be more akin to the machine which can experience colour than any others. However, the first machines we recognize as conscious will likely be ones which exhibit consciousness of the sort Torrance describes, and these machines

[9] While I will not rehearse the arguments for each ethical theory in detail, note that ignoring a person's desires for his life will fail to calculate the utility/disutility generated by particular actions, will treat the person as a means to an end, is certainly not something rational people are likely to consent to from behind a veil of ignorance, and demonstrates a lack of care, compassion, and benevolence. None of these ethical theories will condone simply ignoring the desires of a person, although they will almost certainly allow us to take actions counter to those desires in many cases.

[10] This is one reason why advance directives are important, even if fraught with complications: they allow a person to express her wishes in advance to cover circumstances (such as being in a coma) where she cannot do so directly.

[11] An interesting discussion of the connection between self-awareness and moral standing (or personhood, as she puts it) can be found in Mary Anne Warren's discussion of personhood and abortion (Warren 1973) as well as in Scruton (2006).

[12] It might be that consciousness is also unnecessary for having interests, particularly if we consider an Objective-List view of welfare, as Basl (2012), notes. Hence the category of moral patients may extend slightly further than I argue for here; any machine with interests will count, although I am only arguing here that conscious and self-aware machines have interests.

clearly would have the necessary propositional attitudes in order to have interests.[13] As such, my previous argument stands, and such machines would have moral patiency.[14]

3. Intelligence and Autonomy

Thus far I have argued that having interests is what is necessary for moral standing. Since conscious, self-aware machines have interests, they also are moral patients. In general, however, the question of moral standing for machines is raised in the context of artificial intelligence – would an intelligent machine have moral standing? To provide an answer to this general question, we must ask whether we can assume that intelligent machines are conscious and self-aware. If so, we have addressed the moral standing of all intelligent machines; if not, then further work is necessary to clarify the status of the remaining machines.

To respond to this, we must consider what is meant by an intelligent machine. Shane Legg and Marcus Hutter have gathered many of our informal definitions of intelligence and used them to devise a working account of machine intelligence. (Legg and Hutter 2006a, 2006b, 2007) Informally, their definition of intelligence is "Intelligence measures an agent's ability to achieve goals in a wide range of environments."[15] (Legg and Hutter 2007) One key question that emerges from this definition is who determines the goals of the agent. There are two possibilities: one, the agent's goals are always determined by an outside source or, two, the agent's goals are not always determined by an outside source.

Consider the case where the agent's goals are always established by an outside source. In this case, the goals are communicated to the agent in some fashion, and the agent simply uses its resources to accomplish whatever goals it has been given. For instance, my computer takes actions based on user input and the commands dictated by its programming; its actions are always ultimately determined by a human. Such an agent lacks autonomy.[16] Since the agent lacks self-awareness and lacks the ability to formulate goals for itself, the argument for moral standing does not apply; it will not have a desire for continuation or any wishes as to how to live its life. As such, it is in the same category with chairs and tables mentioned above and lacks moral standing; it is not clear how one could harm or benefit such an entity.[17]

---

[13] I believe that we are more likely to recognize as conscious a machine which has a robust consciousness since that consciousness is more like our own and thus more apt to display behaviors which match up with the conscious behaviors of humans. It is far from clear how we would ever determine that machine had an awareness of colors if that were the full extent of its consciousness. Hence while we may create such limited machines, I suspect we will not realize we have done so.

[14] Marie-des-Neiges Ruffo (2012) would likely object to this conclusion as she believes that machines are not things which are capable of well-being or ill-being because they lack human feelings. I find this unconvincing for two reasons. First, I believe you could create a case which paralleled the congenital analgesia example and argue that it is still wrong to harm such a person even if she lacked emotion. Second, it is not clear to me why she assumes that we will never be able to create machines which have emotions. It is true that we cannot currently do so, but there was a time when everyone was certain a machine would never be able to play chess. This has, of course, proven false; as such, I find our current capabilities to be poor predictors of future ability.

[15] They provide a formal definition (Legg and Hutter 2007), however space does not permit the detailed exposition required to fully explicate this definition.

[16] I am using "autonomy" in the sense typical of ethics, meaning something akin to "being able to make one's decisions free of external influence or control;" the term is (confusingly) used somewhat differently at times in robotics.

[17] Presumably the machine is not sentient, or we could have had a much shorter argument for moral standing; as such, it cannot gain moral rights through an appeal to sentience. One might try to argue that such a being has rationality and thus, on some views of morality at least, must be granted moral standing. I am not convinced this is

Contrariwise, consider the case where the agent's goals are not always determined by an outside source, i.e., where the agent is capable of determining its own goals at least some of the time[18]. In this case, the agent is expressing a basic capacity for autonomy, which implies that these goals must be chosen by the agent itself; they cannot simply be chosen by following an algorithm or program.[19] As such, the agent must be deciding for itself what it desires to do. Once an agent is capable of exhibiting desires, however, we may collapse this into my previous argument concerning moral standing; while the agent's desires may be overridden, they may not simply be ignored.[20]

One could object that this argument is prejudiced by the use of the word "desires" – perhaps the machine is choosing what to do, one might argue, but that does not imply that the machine desires that course of action. Yet, it is not at all clear what such a choice would mean, in this case, since it cannot be determined by an algorithm or program. The machine would need to be making a decision which was in some sense its own; it could not be purely the result of an outside influence or program. Where else would the choice stem from if not from the machine's own wishes? If we have eliminated any external factors or internal compulsion, what remains is the machine's own will.[21]

One point worth noting is that moral questions are not black-and white; both autonomy and moral standing exist on a continuum.[22] The more autonomous the machine, the more duty we will have to respect its wishes; the less autonomy, the more we are permitted to act as its guardian. This is akin to how we treat children and the severely mentally disabled; they are not viewed as capable of making decisions in as many areas as fully-functioning adults, hence we do not see their desires as binding to the same extent. They still have moral standing, of course, in that it is wrong to harm them without just cause. Nevertheless, they are not granted as much governance over the course of their own lives, and we do not view overriding their wishes as comparable to overriding the wishes of other adults. In a similar fashion, a machine with greater

---

the case; while Kant sees morality as shared by rational beings, he makes it clear that the kinds of beings he is discussing have a will – the machines, as I have described them, do not. (Kant 1786/1996) In general, I believe that the rationality criterion for moral standing is more complex than simple intelligence, and machines with bare intelligence will likely not satisfy it.

[18] It is not clear whether such a machine currently exists; I suspect it does not yet, although the evolution of drone technology seems to be heading us in this direction.

[19] While the choices may be influenced by the programming of the machine, human choices are also influenced by upbringing, societal pressure, brain chemistry, and so forth. Since moral theorizing generally views human autonomy as worth preserving despite these factors, machine autonomy likewise has worth.

[20] One might also make the argument that autonomy itself is sufficient for granting something moral standing. If we view autonomy as a good, then the fact that such machines exhibit autonomy suffices to grant them at least some consideration. We may place limits on the expression of their autonomy, just as we do for people, but we likely could not simply ignore it.

[21] Note that this argument is separate from the argument of whether such machines could exist. Ruffo (2012) believes that a machine cannot deliberate; any choice it makes would be a result of programming. As such, she would argue that no machine could determine its own goals. While I am unconvinced, it is not necessary for our present purposes.

[22] A machine which is programmed to learn based on past interactions will be somewhere along this continuum, depending on the complexity of its programming; a simple program will likely result in a machine with little autonomy, but a complex program may approach the situation we have with humans. Since we also learn and adapt as a result of our interactions – following social norms, rules we have been taught, biological imperatives, and so forth – a sufficiently complex set of instructions for a machine may model this; if we consider ourselves to be at least somewhat autonomous, we must consider the machine to be as well.

autonomy likely has more claim on us to respect that autonomy, and it will be a greater moral fault if we ignore its wishes.[23]

In summary, I believe that autonomy implies that the agent has desires. My previous argument fails to apply only to intelligent machines which both lack self-awareness and consciousness and also which are not capable of setting their own goals. Such machines lack moral standing because they have no self-concept and no desires; it is implausible to hold that they could desire existence or have goals for that existence. Determining whether and to what extent a machine is autonomous will likely be difficult, however, and those who oppose granting moral standing to machines might well use this as an excuse to deny their moral worth. This is a dangerous move to make, though, since the long-standing philosophical dilemma of other minds demonstrates that it is also hard to ensure that other people have minds and are not cleverly programmed automata which simply deceive us into thinking they are conscious humans.

The problem of how to determine whether machines are conscious or autonomous is difficult. Torrance (2012) seems fairly optimistic about the prospect; assuming that consciousness is not simply some mysterious fact about the universe, then it likely hinges on facts about the structure of our brains and is exhibited in our behaviour. Hence, in general, we assume that a person is conscious because she acts in particular ways and because she has certain biological similarities to ourselves; if we take ourselves to be conscious, then a creature which acts like us and is built like us seems likely to be as well. Yet, as Basl points out, this is challenging to extend to machines because they are not like us biologically. Even if we build machines with biological components, they will not share the same evolutionary history as we do; hence it is more difficult to argue that their minds have developed similarly (and thus must also have given rise to consciousness.) (Basl 2012) Perhaps as our knowledge of what is physically necessary for consciousness in humans progresses will be able to recreate it in an artificial setting; for now, it leaves us in a bit of a quandary.[24]

In general, it is wise to err on the side of caution – if something acts sufficiently like me in a wide range of situations, then I should extend moral standing to it.[25] Joanna Bryson (2010) has argued that there is danger in being overly generous and extending rights to machines because we may waste energy and resources on entities which are undeserving of them; furthermore, this diverts our attention from the human problems which should be our concern.[26] However, I believe she is too hasty in arguing that we simply can avoid the problem by not designing robots which deserve moral concern. While she is correct that we design and build robots, it should be clear to anyone who interacts with computers or software that we do not always correctly predict

[23] The analogy is somewhat imperfect, since we tend to take children to be beings who will increase in autonomy over time; they have the potential for as much autonomy as fully-functioning adults, whereas we generally are not as optimistic about the prospects of the severely mentally disabled. However, I can see the potential for both sorts of machines: there may be some whose autonomy only ever reaches a low level and others whose autonomy develops over time. Hence the two prongs of this analogy are both useful, since I believe our treatment of those machines ought to parallel our treatment of similar humans.

[24] For that matter, we could likely repeat this argument when addressing the question of whether a machine can have a mind, since again such a machine will not share our evolutionary history and so forth.

[25] Think of this as the moral equivalent of the Turing Test: if the machine's behaviour is indistinguishable from a human's behaviour in most situations, then there is a prima facie case for treating it similarly. This argument is used by Peter Singer (2002) to argue for our assumptions of sentience both in other people and in animals. A similar line of thought has been developed by Rob Sparrow (2004, 2012) in trying to determine when we would view a machine as similar enough to a human to warrant the same moral standing.

[26] This concern has been echoed by Torrance (2012), although he seems more sympathetic to the dangers of mistakenly denying rights to machines which deserve them.

the results of our creations; moreover, there will almost certainly be someone who tries to design a self-aware autonomous machine simply because he can – because it would be interesting.[27] As such, it is overly optimistic to believe we can simply avoid the question in the manner she suggests.

Once we acknowledge that someone is likely to try to create such machines, or perhaps has even done so**,** we cannot ignore the question of appropriate moral standing. At least two pertinent objections must be acknowledged. First, there is the concern that by extending moral standing too widely with respect to machines we might unjustly limit the rights of the creators or purported owners of said machines: if, in fact, those machines are not autonomous or self-aware, then we have denied the property claims of their owners. Second, there is Bryson's concern that we may waste resources by extending rights to machines that are not autonomous. In each of these cases, however, I see the moral fault in being overly conservative is much larger than the risk of being overly generous.

The risk of losing a piece of property is trivial compared to denying moral standing to a being. However, it is much more difficult to dismiss the larger concern that, as a society, we may divert resources inappropriately; we have difficulty using our resources to aid the humans we know have moral standing, and the problem only magnifies when we consider the case of machines. It is certainly reasonable to recognize that our resources are limited and we may not be able to help all persons. Yet surely it is morally repugnant to allow someone to freeze to death because we had diverted our energy to power my toaster. While we may must sometimes make difficult choices in situations where we cannot help all people, even those who are ultimately unaided must be given ethical consideration; we cannot simply ignore them.[28] Yet, of course, there is great uncertainty in determining whether machines are moral persons; how then should we address the worry of diverting resources inappropriately?

I believe the best approach is a probabilistic one. We generally believe that other humans are sufficiently like us in various respects that we see as relevant to having moral status. While the problem of other minds raises the possibility that we are deceived, most of us regard it as relatively unlikely; we view the probability of error as too small to risk denying moral standing based on that possibility. While we cannot directly know what it is like to be another human, we use our best understanding to draw parallels with our own experiences; we then make decisions based on that understanding.

We do the same thing when considering the moral status of non-human animals, albeit with a higher probability of error. Hence we may examine the behaviours of chimpanzees, the brain activity in various animals, and so forth. We then compare this to our criteria for moral standing and ask how likely the entity is to satisfy those criteria; if an animal whimpers when you step on its paw, what is the probability that its whimper is a sign of pain? If an animal appears to exhibit emotions, does that make it psychologically like us? We assign moral statuses depending on our answers to these questions, and we reassign statuses when new data shows us that our previous beliefs were mistaken; this is why we have moved away from using chimpanzees in medical research, for instance. (National Institute of Health 2013)

---

[27] There are already many researchers involved in trying to create intelligent machines, for instance via The Mind Machine Project at MIT. Furthermore, there has been a great deal of discussion about what consciousness or self-awareness in a machine would entail. For a number of optimistic outlooks on the matter see Long and Kelley (2010), O'Regan (2012), Gorbenko et al. (2012).

[28] This is why presumably no matter what decision one makes in the trolley case, one is acting unethically if she fails to consider the humanity of all of the people involves. Simply ignoring the personhood of any of the individuals involved is not an ethical move, no matter how much simpler it would make the scenario.

Our application of moral criteria to beings other than ourselves always rests on a kind of estimation because it is not possible to have first-hand experience of others' situations. In the case of non-human animals, as well as in the case of other humans, we are able to find biological similarities to ourselves. However, we will clearly have some evidence in the machine case as well: the question will then be how likely we believe that its behaviours stem from consciousness and self-awareness as opposed to mere programming. Whether we acknowledge it as a moral patient will depend on our answer; if it seems highly likely, then we are more apt to divert resources to it. If it seems less likely, then we may only divert those resources if they are not needed for others.

We will undoubtedly be mistaken in our estimates at times. A failure to acknowledge the moral standing of machines does not imply that they actually lack moral standing; we are simply being unjust in such cases, as we have frequently been before. I am inclined to be generous about moral standing, however, because history suggests that humans naturally tend to underestimate the moral status of those who are different. We have seen women and children treated as property; even today many victims of human trafficking are still treated this way. Under the auspices of colonialism, entire existing civilizations of people of colour were dismissed as inferior to those of white Europeans. Animals remain a source of contention, despite the fact that they seem to suffer. I believe that we are already very sceptical about the status of others; as such, I am less worried that we will be overly generous to machines and more worried that we will completely ignore their standing. I see the risk of diverting resources inappropriately away from machines as far less likely than the risk of enslaving moral persons simply because they are physically unlike us.[29]

4. Moral Standing and Rights

4.1 Rights of Machines

The moral standing of intelligent autonomous machines is a natural extension of the sentience-based criteria for moral standing.[30] Intelligent, self-aware machines are beings which have interests and therefore have the capacity to be harmed. Hence, they have at a minimum moral claims to self-preservation and autonomy, subject to the usual limits necessary to guarantee the rights of other community members.

It is difficult to specify what moral entitlements said machines will have until we know the nature of those machines. For instance, Kevin Warwick (2012) discusses the possibility of conscious robots with biological brains. If those brains contain a sufficient number of human neurons, then they deserve the same kinds of protections we give to other beings with such neural complexity; we would be committing a moral wrong if we treated them as simply a thing in a laboratory. However, since machines (whether a biological hybrid or not) are physically rather different than humans, some rights will need to be "translated." A basic human right to sustenance will take a rather different form for machines, for instance, since they are unlikely to

[29] Some claim that this argument could be used to extend rights to a foetus. However, I think it clear that a foetus does not *at the time it is a foetus* act like me in a wide range of situations; we weigh the probability of its personhood as less than that of an adult human, although how much less will depend on the individual.

[30] It is probably possible also to defend granting moral standing to such machines on a rationality-based understanding of the moral community, however as I am sympathetic to the criticisms of such theories, I shall not attempt to do so here.

need food and water; they might well have an similar need for access to electricity, however. Similarly, just as humans have a need for medical care of various kinds, intelligent machines might require certain kinds of preventative maintenance or repairs.

Even such basic rights raise issues concerning what it means for these machines to exist or to cease to exist. In order to have a right to self-preservation, we must understand what that means with respect to these beings. At present, I can create a copy of a file which is functionally identical to the original. If I copy it on to two separate computers, we generally say that the same file is on each computer. What happens if this is possible to do with a virtual consciousness? Does the entity survive so long as at least one copy remains? If there are multiple copies, does that mean that there are now multiple copies of the same entity? Or are each separate entities with separate identities? Can we destroy an intelligent machine as long as we have copied all of its files onto another machine? What is life and death to a machine?[31]

Moving beyond basic needs for survival, consider rights on a larger socio-political scale, such as the basic human rights espoused in the United Nations' *Declaration of Human Rights* (1948). It is not immediately obvious how some of these will be handled, such as the claim that everyone has the right to a nationality. For humans, we determine that nationality based on the arbitrary criterion of birthplace (or parental nationality); it is then theoretically possible to change affiliation by undergoing certain processes.[32] One might suggest, therefore, that we could grant machines a starting nationality based on where they were first "switched on."

4.2 Rights of Virtual Entities

This answer is further complicated if we extend moral consideration from machines to entities which are not embodied and have only a virtual presence.[33] My argument could fairly easily be expanded to include these entities, since they could also display autonomy or self-awareness. The main adjustment needed is to devise an understanding of what their existence consists in, since it cannot be linked easily to embodiment. We do not have much experience with non-corporeal existence, hence there are metaphysical questions that would need to be addressed before we can determine how best to understand the rights of these beings.

For instance, the human sense of self is frequently tied to our physical embodiment; we see our bodies as part of who we are, which is why people who undergo procedures such as mastectomies often struggle to see themselves as the same person. (Piot-Ziegler et al. 2010) This strong connection to our bodies makes it hard for us to comprehend what sort of identity a disembodied being would have. Clearly such a being should be able to have an identity, however, since even after an amputation or a mastectomy a person retains some sense of self, even if somewhat altered. As such, a specific embodied form is not a requirement for identity and self-awareness. Similarly, the desires of many people not to be kept alive in persistent

[31] This touches on questions relevant to moral agency as well, since as people have noted (Asaro 2012), having legal responsibility would require us to be able to punish a machine which failed in its legal responsibilities; this requires us to know whether and how it is possible to do so.

[32] I say "theoretically" since, in practice, the change of nationality is fairly difficult; most people are pragmatically limited to the nationality of their birth, regardless of having a human right to change it.

[33] One could object that, speaking precisely, such entities will likely not be wholly virtual. Rather, they may well require the existence of physical objects in the same way that computer viruses require physical machines on which to reside; their existence is not independent of physical objects. However, the identity of the virus or the machine is quite distinct from the physical object(s) they depend on in a way unlike our experience of other identities; if they are embodied, it is in a very different sense than we currently understand.

vegetative states highlights the fact that for many the important component of identity is not the body. Together, this implies that a virtual entity could have an identity. Yet clearly this sort of entity will complicate questions such as nationality: how do you attach a nationality to something which does not have a physical presence per se? Is there any benefit to trying to do so? What would it mean if they existed outside the current borders of our political structures?

One possible avenue for investigation is to consider how we treat the moral status of other non-biological entities, such as corporations; they too have an existence which is not directly tied to a particular physical instantiation. A number of philosophers (Wallach and Allen 2009, Asaro 2012) have noted that we have granted legal rights and responsibilities to corporations, effectively treating them like persons. Furthermore, corporations can commit moral wrongs, such as outsourcing garment production to places which lack reasonable safety precautions for workers; corporations which do this are unethically placing profits ahead of human life. These kinds of issues particularly occur with multi-national corporations; we are struggling with how to apply the notions of rights and responsibilities to entities which are not tied to a single location and physical entity. While machines will differ from corporations in a variety of ways, the parallel highlights the fact that we have made this kind of extension of morality before; there may be no easy answers, but we are not left entirely without precedent.[34]

In addition to translating current human rights into forms which are more applicable to machines, it will likely be necessary to consider new problems which these machines generate. For instance, at the moment, it is not possible to replicate the contents of my mind. As such, my identity is fairly solidly unique and not in need of protection. However, as artificial brains become possible, we must ask whether a person has some kind of uniqueness rights; if we copy a virtual entity on to another machine without its permission, have we wronged it?[35] These questions will be necessary to consider as we move forward with artificial intelligence. Hence while it is clear that conscious, self-aware machines have moral standing, it is much more difficult to say exactly what that standing grants them; much depends on how our technologies evolve.

5. Conclusion

I have argued that the sentience criterion for moral standing is, in fact, insufficient to cover all humans; it does not explain what is morally wrong about one's action in the congenital analgesia case. Rather than seeing sentience as necessary for moral standing, therefore, I have suggested a move to an interest-based account: if a being has interests, then it is wrong to ignore those interests or to harm them in the absence of some suitable overriding reason. This view of moral patiency, however, may be extended to machines. If a machine has interests, then it may be harmed or benefitted; it deserves some moral consideration.

Furthermore, I have argued that there are several ways that a being may have interests in addition to being sentient. A self-aware conscious being will have interests; so will an autonomous intelligent machine. In general, if a being is capable of desiring its own continuance and of forming wishes about its future, then we have some prima facie obligation to respect those desires.[36] Determining the details of machines' moral standing is difficult, particularly since the

[34] There is, of course, debate about whether this is a good precedent to have set. The point remains, however, that we have dealt with non-human persons in the law before; it is not entirely new territory.
[35] This is similar to questions raised by cloning a person without permission.
[36] As with any other member of the moral community, those rights may be overridden if necessary.

relevant machines do not yet exist (or at least are not acknowledged to exist); some moral theorizing may need to wait until we have a better idea of what they are like. However, we have some precedent for thinking about the moral standing of non-biological entities by considering the moral status of corporations.

The battle for recognition of machines' moral standing will not be easy. We do not acknowledge the claims of others readily, even when the only difference between ourselves and those people is skin colour or gender; this difficulty will be magnified for intelligent machines. One key problem is the need for others to acknowledge the autonomy and/or consciousness of those machines. Philosophers have been arguing over the problem of other minds for millennia with respect to humans; the problem will likely magnify for machines, since we do not have a clear set of criteria that all will accept as sufficient for consciousness or autonomy.[37] The risk of attributing incorrect moral standing to machines is one which will likely plague discussions of machine ethics for some time to come.

Although I acknowledge that we make mistakes in our attribution, I believe that we are more likely to err on the side of conservatism than that of excess. Not only do we have a long history of doing so to other humans, given our past experiences with colonialism, but there will likely be intense financial pressure not to recognize machines as moral persons. We depend upon machines to do many tasks for us, and we do not currently pay machines or worry about their needs (beyond perhaps basic maintenance). One of the rights enshrined by the United Nations (1948) is the right to remuneration for work, meaning that the financial pressure to avoid recognizing any moral standing for intelligent machines will likely rival the push to avoid acknowledging African-Americans as full persons in the Confederate South. However, we cannot ethically deny someone moral standing simply because it is convenient.

The time to start thinking about these issues is now, before we are quite at the position of having such beings to contend with. If we do not face these questions as a society, we will likely perpetrate injustices on many who, in fact, deserve to be regarded as members of the moral community. I urge moral generosity when considering the moral claims of machines; we need to counter our legacy of sluggishness in recognizing as moral persons those who are physically unlike us.

References

Asaro, P. (2012) A Body to Kick, but Still No Soul to Damn. In P. Lin, K. Abney & G.A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT P, Cambridge, USA.

Basl, J. (2012) Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

Bentham, J. (1996) *An Introduction to the Principles of Morals and Legislation*. J.H. Burns and H.L.A. Hart (Eds.) Oxford UP, New York, USA.

Bringsjord, S. (2010) Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness. *Metaphilosophy*, 41, 292-312.

[37] See Floridi's presentation of this conundrum (Floridi 2005) and an attempt to devise a test for self-consciousness in response (Bringsjord 2010).

Bryson, J. (2010) Robots Should be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues.* John Benjamins, USA.

Code, L. (1991) Is the Sex of the Knower Epistemologically Significant? In: *What Can She Know?: Feminist Theory and the Construction of Knowledge.* Cornell UP, Ithaca, USA, 1-26.

Floridi, L. (2005) Consciousness, Agents and the Knowledge Game. *Minds and Machines*, 15, 415-444.

Gorbenko, A., Popov, V., & Sheka, A. (2012) Robot Self Awareness: Exploration of Internal States. Applied Mathematical Sciences, 6: 675-688.

Gunkel, D. J. (2012) A Vindication of the Rights of Machines. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

Kant, I. (1996) Groundwork of The Metaphysics of Morals. In Gregor, M. (Ed.), *Practical Philosophy*. Cambridge UP, Cambridge, U.K.

Legg, S. & Hutter, M. (2006a) A Collection of Definitions of Intelligence. In Goertzel, B. (Ed.), *Proc. 1st Annual artificial general intelligence workshop.*

Legg, S. & Hutter, M. (2006b) A Formal Measure of Machine Intelligence. In *Proc. Annual machine learning conference of Belgium and The Netherlands*. Ghent, Belgium.

Legg, S. & Hutter, M. (2007) Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17, 391-444.

Long, L. N. & Kelley, T.D. (2010) Review of Consciousness and the Possibility of Conscious Robots. Journal of Aerospace Computing, Information, and Communication, 7: 68-84.

Mill, J.S. (1993) *On Liberty and Utilitarianism*. Bantam, NY, USA.

Mills, C. (1999) *The Racial Contract*. Cornell UP, Ithaca, USA.

National Institute of Health. (2013) Council of Councils Working Group on the Use of Chimpanzees in NIH-Supported Research Report. http://dpcpsi.nih.gov/council/pdf/FNL_Report_WG_Chimpanzees.pdf. Accessed 6 March 2013.

O'Regan, J. K. (2012) How to Build a Robot that is Conscious and Feels. *Minds and Machines*, 22: 117-136.

Piot-Ziegler, C. et al. (2010) Mastectomy, body deconstruction, and impact on identity: A qualitative study. *British Journal of Health Psychology*, 15: 479–510.

Ruffo, M. (2012) The robot, a stranger to ethics. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

Scruton, R. (2006) *Animal Rights and Wrongs.* Continuum, London, U.K.

Singer, P. (2002) *Animal Liberation*. Ecco, USA.

Sparrow, R. (2004) The Turing Triage Test. *Ethics and Information Technology*, 6, 203-213.

Sparrow, R. (2012) Can Machines Be People? In P. Lin, K. Abney & G.A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT P, Cambridge, USA.

Taylor, A. (1996) Nasty, brutish, and short: The illiberal intuition that animals don't count. *The Journal of Value Inquiry*, 30: 265-277.

Torrance, S. (2012) The centrality of machine consciousness to machine ethics: Between realism and social-relationism. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

United Nations. (1948) The Universal Declaration of Human Rights. http://www.un.org/en/documents/udhr/. Accessed 3 January 2013.

Wallach, W. & Allen, C. (2009) *Moral Machines: Teaching Robots Right from Wrong*. Oxford UP, Oxford, U.K.

Warren, M. A. (1973) On the Moral and Legal Status of Abortion. *Monist*, 57, 43-61

Warwick, K. (2012) Robots with Biological Brains. In P. Lin, K. Abney & G.A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT P, Cambridge, USA.

Zack, N. (2002) *The Philosophy of Science and Race*. Routledge, New York, USA.