Abstract:

This thesis deals with moral status and its potential application to machines. It introduces an account of moral status and defends the claim that with the correct features a machine can have moral status. The thesis also discusses some issues involved in recognizing the non-apparent features of a machine and how we might overcome them.

THE FLORIDA STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES

AN ACCOUNT OF MORAL STATUS FOR MACHINES

By

RANDY BRUNO-PIVERGER

A Thesis submitted to the
Department of Philosophy
in partial fulfillment of the requirements for graduation with
Honors in the Major

Degree Awarded:
Summer Semester, 2018

The members of the Defense Committee approve the thesis of Randy Bruno-Piverger defended on December 7th, 2017

Dr. Simon Căbulea May
Thesis Director

Dr. Mark LeBar
Committee Member

Dr. David Gaitros
Outside Committee Member

Signatures are on file with the Honors Program.

Contents

A Case Involving the Problem

In Ridley Scott's *Blade Runner* Harrison Ford plays Deckard, a former blade runner, who specializes in hunting down and "retiring"[1] the human-like androids known as replicants. Replicants, at least superficially, appear as the main antagonists of the film. They are developed artificially and programmed with awareness of their artificiality. They are also designed to continually evolve intellectually. To prevent the androids from fully realizing their evolutionary potential, and possibly revolting against their human creators, each replicant is designed to have a four-year lifespan. The story begins with Deckard being called into work by his former boss. Deckard is briefed on a group of rogue replicants that escaped from their labor camps and took over a space vessel, killing its crew and using the ship to land on Earth. Deckard is hired to retire the replicants before their presence on the planet is made public. With considerable difficulty Deckard is able to retire all but one of his targets. The final replicant, Batty, faces off against Deckard and defeats him. Now hanging on the edge of a tall building for dear life, Deckard is rescued by the same replicant he aimed to kill. The narrative reveals that the replicants only returned to Earth, to track down their creators and find a way to extend their lifespans. Batty, now at the end of his, divulges his frustrations to Deckard. He recalls some of the most

[1] Scott, Ridley, director. Blade Runner. Warner Brothers, 1982. - The killing or decommissioning (depending on one's interpretation) of replicants who's mental development represents a threat to humanity.

memorable moments of his life, moments of beauty and humility, and he laments at the thought that all his deeds and all his memories would be lost in time like "tears in the rain." [2] Almost immediately after his monologue Batty dies, bringing an end Deckard's mission.

This scene is designed to convince us that Batty is already much more than just an android. The most straightforward example of this is Batty's ability to show compassion to Deckard, saving his life even after witnessing Deckard kill two of his comrades. More interesting is Batty's ability to reflect on his life and share feelings of dissatisfaction and sadness now that he is faced with its brevity. Moreover he is dealing with how to come to grips with death itself which he describes as having the collection of moments that make up your existence fade into obscurity. Most cannot help but feel saddened that Batty has died by the end of the scene but if Batty is nothing more than an android are our feelings of sadness in response to his untimely death appropriate? And if we do feel remorse over the death of Batty, is it because he is obviously someone who has been morally wronged? Or, does this scene represent some kind of trickery performed on us by the movie's creator?

[2] Scott, Ridley, director. Blade Runner. Warner Brothers, 1982.

Thesis

The chief goal of this paper is to substantiate the claim that a machine can hold moral status if and only if it has the capacity for "well-being." [3] To argue this, I will employ an account of moral obligation posited by Alasdair Cochrane which maintains that entities with the capacity for "well-being" have moral status.[4] This account refutes the claim that all living organisms have well-being and argues that "phenomenal consciousness" (sentience, from here onward) makes it possible for beings to have lives that go well or badly for themselves.[5] It also requires that sentience, in this sense, is not only the capacity for consciousness but also the ability to feel good or bad about one's condition—evaluative sentience, as it were. Finally, the paper will discuss how we can recognize the necessary features for moral status in machines.

In Section I of this paper, I will begin by defining moral status for the purposes of this discussion. Section II will present the argument supporting the claim that machines can have moral status. In section III, I will discuss my method for recognizing the necessary features for moral status in machines, then review two alternative methods and provide critiques for each. The final section will respond to potential rebuttal arguments.

[3] Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." p. 20.
[4] Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." p. 20.
[5] Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." p. 20.

I. Moral Status

A. Defining Moral Status

To have moral status is to be morally considerable.[6] Being morally considerable involves being an entity which moral agents (i.e. individuals capable of making decision based on some notion of right and wrong) have or can have moral obligations towards.[7] A moral obligation should be thought of as a duty that is owed to an entity, such that if that duty was violated the entity would be harmed. Entities that are morally considerable are also frequently referred to as moral patients (a term I will continue to use in this paper). Generally, a moral patient must be treated with particular care. The moral agent must consider the needs, interests, and well-being of moral patients.[8] The importance of doing so is twofold: by respecting the needs and well-being of others the moral agent will benefit other persons and possibly benefit herself, but more importantly, her actions recognize the moral significance of the needs of the entity.[9] Asserting one's moral status not only supposes some sort of moral obligations between you (as a moral patient) and other moral agents, it also affirms an importance of your needs as a moral patient, such that the actions of others are restricted to some degree based on those needs. This inherent importance of moral patients' needs serves as the foundation for moral status and moral

[6] Warren, Mary Anne. "Moral Status: Obligations to Persons and Other Living Things."p3
[7] Warren, Mary Anne. "Moral Status: Obligations to Persons and Other Living Things."p3
[8] Warren, Mary Anne. "Moral Status: Obligations to Persons and Other Living Things."p3
[9] Warren, Mary Anne. "Moral Status: Obligations to Persons and Other Living Things."p3

consideration. An example of a moral agent is a typical human being, as we are capable of making moral decisions. Humans also serve as a good example of moral patients because we are frequently the object of moral obligations. A host of other entities are also often considered moral patients (e.g. Dogs, cats, dolphins etc.).

II. Moral Status in Machines

A. Central Claim

The central claim of this work is that machines can hold moral status if and only if they can have the capacity for 'well-being.' Moreover, we can understand our moral obligations to other entities, including machines, by understanding harm in terms of well-being. The following are essential considerations for these claims.

B. Consideration I: Well-Being and Moral Status

As already mentioned, my argument relies on the consideration of well-being in a given entity. For the purposes of this paper 'well-being' should be understood to refer to an entity's capacity to have a life that can go well or badly for that entity.[10]

[10] Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." p. 20.

The next part of this consideration is the claim that sentience is a necessary and sufficient condition for the capacity of well-being.[11] This is a necessary condition because an entity could not possibly feel or perceive itself worse off or better off without the general ability to feel or perceive to some degree (to have sentience, to a degree). It is also a sufficient condition for the capacity of well-being because the ability to perceive, feel or experience the world in a subjective manner also suggests the ability to feel better or worse about the situation an entity finds itself in at any given moment.

These capabilities alone, if possessed by a machine, would be a clear indication of moral status. One way to ascertain whether a machine is in possession of these capabilities, is discussed later, in Section III.

C. Consideration II: Needs and Interests

Martha Nussbaum defines "needs" as "central human functional capabilities." [12] She outlines ten essential functional capabilities: life, bodily health, bodily integrity, imagination and senses, emotion, practical reason, affiliation, relationship with other species, recreation and control over one's environment.[13] She argues that one must be capable of achieving all these functions in order for a human's life to go well. One's well-being rests entirely on one's needs and because needs are essential to living a good life, moral consideration requires us to respect the needs of other moral patients, and consider them before making decisions. For machines, I

[11] Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." p. 20.
[12] Martha C. Nussbaum, Women and Human Development: the Capabilities Approach, (Cambridge: Cambridge University Press, 2000), pp. 78-80.
[13] Martha C. Nussbaum, Women and Human Development: the Capabilities Approach, (Cambridge: Cambridge University Press, 2000), pp. 78-80.

suggest we take a similar approach and define their needs based on their functional capabilities. In sentient machines, we might insist that their life, bodily health, bodily integrity, imagination, senses, and emotion be considered needs, and not be unreasonably molested. These specifications arise naturally from what it means to be sentient and capable of perceiving the world around you.

"Interests" are defined by an account from Joel Feinberg.[14] On this view, to have an interest x means having a stake in x. Having a stake in something, x for instance, means being in the position to gain or lose depending on the condition of x. Gaining or losing on this view has to do with the status of one's well-being. Being better off due to the state of an interest would correspond to a gain in well-being and being worse off due to the state of an interest would correspond to loss in well-being. Interests connect to our well-being through our needs. Interests arise from our needs, instrumentally, and ultimately exist to further our well-being. Altogether, when we say that one has an interest, we mean that one's well-being is affected by the state of some endeavor and consequently the quality of one's life is related directly to the condition of that endeavor. For example, a surgeon has an interest in keeping her hands in a healthy condition, more so perhaps than your average individual. This is an interest because her hands help her in sustaining her way of life. They allow her to continue her work which pays for her housing, her food, her medicine etc. and other items that satisfy the essential needs of the doctor and ultimately benefit her well-being. Assaulting the surgeon and wounding her hands would therefore represent a significant violation of her interests as well as a disregard for have needs as a moral patient. Interests also inform our moral obligations. For example, a dog has a clear

[14] Feinberg, Joel, "The Rights of Animals and Unborn Generations" p. 33-34.

interest not to be hit by a car while crossing the street. Getting hit would, at best, wound the animal and put her in a state of pain. At worst, it would kill her. Both these consequences represent clear losses in well-being as well as a disregard for the needs of the animal. In showing respect for the needs of the dog we must recognize that we now have an obligation not to hit that dog with our car.

Since the capacity to have interests imply the capacity to have a well-being it can also be said that a machine with interests (by definition) must also have moral status.

D. Consideration III: Harm

This consideration is also closely related to well-being. We will use another account from Feinberg. He defines harm as a "thwarting, setting back or defeating of an interest."[15] Given the previous consideration we can take this to mean that harm happens when an interest is foiled or set back resulting in a loss of well-being. There is at least one problem with defining harm in this way – namely, its lack of complexity. Stephan Wilkinson offers a rather crude hypothetical situation to illustrate the deficiency of this definition. It involves a psychotherapist who sexually exploits his patients.[16] Despite what we already intuitively know about this scenario, that taking advantage of your patients as a doctor cannot be good, it is altogether possible that one of the man's patients will leave his office with their well-being in better condition than before they arrived. A similar example could be made of a fraudulent tax accountant. Let's say an accountant offers to save a client thousands of dollars on his taxes, and secure for him thousands of dollars

[15] Feinberg, Joel, "The Rights of Animals and Unborn Generations" p. 33-34.
[16] Wilkinson, Stephen, "Bodies for Sale: Ethics and Exploitation in the Human Body Trade" p. 60.

in tax refund money. This accountant ends up taking more money from the tax return than what was agreed upon in his contract, but he does succeed in saving the man thousands of dollars and securing him a commendable amount of money in the form of a hefty tax refund. The rub is that despite the man being better off than he was before seeing the accountant it still seems clear that he is being wronged here. In response to this, Cochrane makes an amendment to Feinberg's definition of harm; it outlines not only what actions hurt our well-being, but also what actions leave our well-being worse off than it should be.[17] What "should be" ought to be thought of as the possible actions available to the moral agent that creates optimal results for their moral patients without causing unjustified harm to themselves or their moral patient. In this case, the course of action that would have been most beneficial to both the accountant and to man would be if the account stuck to terms outlines in his contract. Of course the accountant would not have nearly as much money at the end of the day and his well-being would have gained less when compared to the well-being of an accountant who has stolen from his client. However, theft on the part of the accountant represents an unjustifiable kind of harm inflicted upon his moral patient, the action also disrespects the importance of his client's needs. There are scenarios, however, where causing some harm can be reasonable and morally justifiable. In situations where causing some minor amount of harm will prevent a more serious harm from occurring, harm can be morally acceptable. For example, open heart surgery inflicts a considerable amount of harm to the body but if the only alternative is death, than the option is morally permissible. Moreover, if a decision can cause unreasonable amounts of harm to the moral agent, even if it

[17] Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." p. 24.

provides the highest benefit to the well-being of her moral patient, it is reasonable and morally permissible that this moral agent would not take such an action.

III. Apparent Features and Assumed Features

The preceding text outlined necessary features and considerations for judging whether or not a machine has moral status. It is at this point that I must turn the focus of the paper from necessary requirements for moral status to how it is we observe these requirements in machines. This is a particularly contentious question but to answer it in short, it is simply the process of substantiating or overcoming one's doubts about a machine such that one is sufficiently confident in whether or not an entity has moral status. This is done by reviewing the "apparent features"[18] of the machine to the point the observer is confident enough to decide what non-apparent features they believe the machine does or does not have. Here I offer my thoughts on how determining moral status for machines could potentially be organized. Afterward, I review two alternative accounts of moral consideration for non-human entities.

"Apparent features" are described by Mark Coeckelbergh as the features or abilities of an entity that are used as criteria to base our moral considerations upon.[19] Apparent features are "features-as-experienced-by-us."[20] They include all objectively observable things we can use to characterize the non-apparent features of an entity. Non-apparent features are assumed features

[18] Coeckelbergh, Mark. "Robot rights?…"
[19] Coeckelbergh, Mark. "Robot rights?…"
[20] "Criteria for Recognizing Sentience." Animal Ethics.

that we cannot objectively observe. Features of this sort would include sentience, rationality, and self-awareness. For moral status the only non-apparent feature we are burdened with is sentience. Sentience is the capability to perceive or feel.

To deduce whether or not a machine has sentience we must investigate the apparent features that relate to this non-apparent feature. We could start with the machine's physiological or mechanical make up. A nerve structure would be strong indication that the machine is capable of experiencing some level of sensation, possibly pain. This would be clear indication of sentience. Next we ought to inspect the evolutionary course of the machine.[21] Namely, the machine's ability to develop, adapt and modify its own internal capabilities. If the machine does not have already have apparent features that suggest sentience, that doesn't mean it cannot develop them. If the machine is the sort capable of independent growth then this is an apparent feature that is important to pay attention to. Evolution is essentially the story of every sentient being on our planet. The capacity to feel arises in evolutionary history in connection with the necessity to behave in one way over another. [22] This could very well remain true for machines, as well. That established, behavior that is very "plastic", i.e. complex and highly adaptable to circumstance, is also a strong indication of this evolutionary potential.[23] The final relevant apparent features are general behavior. The most obvious evidence of sentience among these features is the evidence of suffering or enjoyment. For a machine we might not be able to observe a smile or frown indicating happiness or sadness, but we may be able to observe a machines understanding of the beneficial and harmful aspects of the environment it resides in.[24]

[21] "Criteria for Recognizing Sentience." Animal Ethics.
[22] "Criteria for Recognizing Sentience." Animal Ethics.
[23] "Criteria for Recognizing Sentience." Animal Ethics.
[24] "Criteria for Recognizing Sentience." Animal Ethics.

For example, let's say a mechanized dog wanders to the basement stairs of its house and slips and falls half way down the stairs. The next day the dog avoids the stairs completely, apparently in recognition of the potential harm of falling down the stairs again. Another example could involve the dog's master playing with her in the same location every day. If one day he begins notice that the dog learned to wait for him in that location ahead of time, perhaps associating that spot with the reward of his attention, it would be a strong sign that the dog is sensitive to beneficial and harmful states, indicating a clear sign of sentience. Each of these examples give apparent features that are closely linked with the non-apparent feature of sentience and any number of them could give an observer strong reason to believe their machine is sentient.

But, there is still a problem here. What if the programming of the machine is built to emulate a sentient creature? What if it is programmed by the best software engineers in the world, making it impossible to distinguish it from a genuinely sentient creature? I would respond with this: if there is no distinguishing between the two, then the question becomes practically meaningless. A creature programmed to perfectly to emulate a sentient creature could very well be a sentient creature (if not by definition, then certainly through observation, though it is difficult to see how we could distinguish between the two in a real world setting). Of course this is why it is important that we investigate these machine thoroughly. Not just by monitoring their behavior, but also by observing their mechanical make up and bodily capabilities. If after all is said and done the one cannot find any reason to not believe that the machine is a moral patient, it is likely in his best interests to treat this entity as a moral patient and avoid potentially harming an entity with moral status.

The fear that the phenomenal feature of moral patients (i.e. sentience), will never be perfectly reproduced in a machine or for that matter, reproduced such that the machine's

behavior alone could convince us of its moral status is not at all unfounded. As it stands today there are no clear metrics for testing phenomenological experiences within machines. So for now we have no way of knowing if a machine has sentience or a self-awareness akin to that of a human. Even if we somehow build an android that did have perfect replicas of those features, we have no way of confirming their existence to such a degree. But this reality should not deter us from continuing to try and find suitable tests for capabilities like intelligence and moral status. Mankind has arrived at a technical paradigm that makes questions like these worth thinking about.

It is also important to note have a shared biological background. Not only are we all organic entities we are of the same species. When describing our phenomenological features we naturally assume that they are qualities we share with other normal humans. This assumption is based roughly on the apparent features of those around us, and also on scientific evidence supporting the idea that humans share very a similar genetic make-up. For example, we observe that Vince is human, we might also note that he appears relatively normal (has no apparent deficiencies), from there, we might naturally assume that Vince has a self-awareness very similar to our own. We make this assumption despite being unable to physically prove that Vince's self-awareness exists in the same way ours does, if exists at all. Granted we may in fact have very strong reasons to believe that Vince is self-aware, e.g. perhaps we've had conversations with him in which he was able to reflect on his past and also tell us his plans for the future – giving you reason to believe that he is able to perceive himself as existing through time, a commonly cited requisite of self-awareness. No matter what evidence we have to support the claim, our beliefs with respect to the non-apparent features Vince and I share are bolstered by the fact that he is human. Now, again this fact may give us stronger reason than most to believe that he shares this

phenomenal quality of self-awareness but what is troublesome about this observation is that when we don't share qualities like species or even basic biological make-up with the object of our observation what we tend to assume is the opposite, that we are unique and that it would take a very special sort of entity to share in our human features. This biased attitude only makes it more difficult to judge non-apparent features.

A. Alternative Account I

Computer Scientists like Alan Turing realized the problem with proving the existence of a phenomenal feature, but also understood that computer technology had the potential to someday become intelligent.[25] Turing does not bother with ideas of measuring computer intelligence through physical means. Instead he created an aptitude test that simply required a machine to convincingly emulate a human in conversation. If the machine could keep its tester from realizing it was a machine then the computer should be considered intelligent. The proposal of this "Turing Test" has had a long lasting effect on the discussion surrounding machine intelligence and how to tell when machines achieve intelligence or self-awareness. [26] I mention Turing because I believe his style of intelligence testing has use for theorists concerned with discerning whether or not a machine has moral status. Robert Sparrow is another theorist who has used the model provided by Turing to develop his own test on the moral status of machines.

[25] Turing, A. M. "I.—Computing Machinery And Intelligence."
[26] Turing, A. M. "I.—Computing Machinery And Intelligence."

Sparrow argues that passing the Turing test is a more than sufficient burden to prove that a machine is intelligent. He believes that the test also demonstrates a machine is self-conscious, and has the ability to form projects and hold ambitions.[27] Additionally, he argues that if a machine can talk like a human it would need the capability to report on its internal states and its past. Ideally, this would demonstrate the machine's self-awareness. Similarly, expressions of contentment, sadness, anger and joy could be demonstrated via conversation. Sparrow relies on the idea that a machine not capable of these things could never pass the Turing Test. Sparrow also proposes a test for recognizing when a machine ought to have moral standing. It is a moral dilemma involving three parties, where the first must choose to save one but at the cost of the life of the other. With all things equal, if we can replace one of the latter two parties with a machine without compromising the difficulty of the moral dilemma, then such a machine must be worthy of moral standing.[28] In other words, when we find that a machine's existence is just as difficult to sacrifice as a human life, then that machine must also be due at least equal moral consideration to that of a human. He calls this the Turing Triage Test. Sparrow claims that the essential capability necessary for moral standing is the capacity to experience pain and pleasure, as it provides at least superficial grounds for moral concern; the basis here is preventing harm.[29] The extent to which this moral concern ought to apply to a being increases depending on how conscious it is of itself as existing across time, its ability to have personal projects and its rational capabilities.[30]

Sparrow's method is rather similar to the one proposed in this paper. Both arguments rely strongly on the observations of an outside entity in deciding whether or not a machine has moral

[27] Sparrow, Robert. "The Turing Triage Test."
[28] Sparrow, Robert. "The Turing Triage Test."
[29] Sparrow, Robert. "The Turing Triage Test."
[30] Sparrow, Robert. "The Turing Triage Test."

status. But a significant difference exists in what comes after that observation. Sparrow uses the reaction of a moral agent to decide whether or not some machine already has moral status, and from there assumes other non-apparent capabilities must reside within that machine (i.e. sentience, self-awareness, etc.). This paper's method observes several apparent behaviors in order to ultimately make an accurate inference as to the non-apparent capabilities the entity has. If the appropriate capability is (or is not) believed to be held by an entity, we can then decide whether or not a machine has moral status.

Fundamentally, I cannot say that Sparrow's method is any less effective than my own at judging whether a machine has moral status. What can be said of his method, however, is that it takes just as much risk as my own method in determining which machines has moral status. Tests that rely on human observation have the disadvantage of assuming features within the machine that are impossible to confirm objectively. This disadvantage is practically insurmountable given the nature of those features.

B. Alternative Account II

Mark Coeckelbergh argues that there is approach to moral consideration of robots, animals, and humans that can help us answer this question. This method presents a "social-relational" justification of moral consideration.[31] It asks that we recognize the experiences we take part in with an entity, x, in the context of a human-x relation that exists within a wider social

[31] Coeckelbergh, Mark. "Robot rights?…"

structure.[32] This emphasis on outward influences in human behavior and experiences characterizes this account; Coeckelbergh seems to be advancing a sort of moral relativist/skeptic theory. The method has four basic tenets.

First, moral consideration must be understood as "extrinsic" to the entity in question; Moral consideration is *ascribed to* entities in social relations within a social context, by other entities.[33] Second, the features or abilities of an entity are used as criteria to base our moral considerations upon. In this case, however, we refer to them as "apparent features," that is to say, "features-as-experienced-by-us."[34] Third, the interactions involving an entity are context-dependent, in that, they require paying attention to the ways in which entities in various social contexts and social relations are granted moral consideration. These experiences are also subject-dependent: they require us to recognize that we can only have knowledge of objects as they appear to us.[35] This implies also that that our observations of the world are to some extent internalizations and that there is no observer-independent reality or "thing-in-itself." This argument is a contradiction to direct arguments for moral standing which assume that certain entities have inherent moral standing in virtue of some internal capability which can be readily recognized by others (e.g. rationality, in the case of Immanuel Kant), or an inalienable right imbued in them by a creator (John Locke).[36] Together the subject-object dependency implies that moral significance arises from the relation between the object and the subject. Finally, we must also view the subject-object relation as being continually shaped in social relations. Here we must recognize that the interactions between subject and object exist prior to the moral

[32] Coeckelbergh, Mark. "Robot rights?…"
[33] Coeckelbergh, Mark. "Robot rights?…"
[34] Coeckelbergh, Mark. "Robot rights?…"
[35] Coeckelbergh, Mark. "Robot rights?…"
[36] Coeckelbergh, Mark. "Robot rights?…"

arguments we engage in, but also those interactions emerge subsequent to some social context.[37] Simply put, there is a social context which greatly affects our moral considerations and which precedes the thoughts we have about how those considerations ought to be governed. Moreover, that social context continually changes. Coeckelbergh points this out to remind us that moral considerations and moral judgements are subject to change. Not only can we see differences in moral considerations through time but also among different cultural spaces—which is yet another aspect of social context.[38]

When trying to determine what degree of moral consideration is appropriate for a particular entity, the social-relational method asks that we take note of the apparent features we notice in an entity, then consider how such an entity might be treated rightly in the social context that we are a part of. We do this by comparing how this entity is similar or different to other entities who share some or all of its apparent features within that social context. We can then make comparisons on how those entities are morally considered and ideally, come to understand what is necessary for certain degrees of moral consideration in certain social contexts.

For example:

> How should one manage one's moral considerations to one's dog? Let's say for the sake of this example the dog suffers from separation anxiety. The owner might ask: What is the nature of human-pet relations that already exist in my social context? Given my dog's particular apparent features how do I modify my moral considerations to better

[37] Coeckelbergh, Mark. "Robot rights?…"
[38] Coeckelbergh, Mark. "Robot rights?…"

> suit the unique traits of my animal? We can observe, that a dog experiences pain and pleasure and thus, obviously has interests. One clear interest our dog has is being near its master and so leaving the dog alone for hours at a time does seem to be a disruption to the dog's interests. We could even go as far as to say that during this time the dog seems to be suffering for it. We also understand that human-pet relations in our social context show us (our society in general) to be invested in the health of our animals, such that we are willing to go out of our way (to some extent at least) to protect their wellbeing. Since we are specifically troubled by their dog's sociable nature and worried about leaving them alone for extended periods of time, we would want to compare our animal with other entities who share its apparent features within our social context. Immediately we might think of humans, who are also social animals. While clearly humans distinguish themselves from dogs via their superior cognitive abilities, we are still vulnerable to the effects of isolation. It seems obvious but worth mentioning that humans need human-to-human interactions to sustain a healthy life. Given that dogs, much like ourselves, are socially sensitive creatures and the dog clearly does not enjoy being left alone it seems reasonable to adopt the general rule, that you will not leave him alone for more than x hours and when you must leave him you will hire a dog sitter.

This process shares at least one thing with Sparrow's test, in that it asks that we compare our impression of a machine's apparent features to a human's, with the goal of making a very serious moral decision. That decision will then, after the fact, reveal to us the relevant apparent factors for offering moral consideration to a machine—considerations comparable to that of a human. Coeckelbergh's method, however, is more flexible. We can allocate general rules that we personally see fitting according to our own conception of the social context and the apparent

features of our object. It also allows us to allocate considerations of varying degrees based on what we feel is appropriate.

One weakness this method avoids is that of prejudice or snobbery on the part of the subject. Sparrow addresses this by expressing a need for quality control with respect to the subjects of each test. "Idiosyncratic" individuals or those with less than average reasoning ability should not be considered proper subjects for the test.[39] This is problematic for a number of reasons. For one, the idea of implementing quality control on moral consideration is worrying because it suggests that there is a proper moral consideration that can be qualitatively and objectively accessed. The social-relational method rejects this idea as moral consideration exists in spaces between people, in relations, as it were, and arise from subject impressions of other social relations, and therefore have no intrinsically right or wrong values attached to them. This is supported by Coeckelbergh's claims on apparent features and the internalization of our observations.[40] On this view, any moral consideration is ultimately the result of external social influences leveraged against an internalized impression of an object or more specifically, an internalized impression of how we ought to interact with an object. With this established we should now ask: What justification do we have to discount another individual's conception of moral consideration? The question "How do we go about measuring reasonableness in the first place?' should follow immediately after. Envisioning a test of reason that is not somehow also inspired by one's own internalized conceptions seems just as troubling as the idea that there are some moral considerations out there that are objectively more reasonable than others.

[39] Sparrow, Robert. "The Turing Triage Test."
[40] Coeckelbergh, Mark. "Robot rights?…"

Conversely, Coeckelbergh's method does differ in many ways from Sparrow's and my own; namely, in his methodology and underlying assumptions. His argument claims that moral status does not exists as-a-matter-of-fact, but is conferred upon an entity given the correct set of apparent features within the correct social context. It also stipulates that nothing objective exists beyond what we can observe. He uses this claim to craft a method that roughly bases its calculations on what behavior is agreeable in a given social context, with a specific kind of entity. Sparrow's method (as well my own method) work to determine the existence of a non-apparent feature. This method denies the existence of that feature (as an internal, non-apparent capacity) completely and instead offers us an open ended guide on when it is appropriate to confer moral status to (as opposed to recognizing it in) another entity.

One obvious issue with this view is that is does not recognize objective standards of right or wrong. If we were to apply this method to our Blade Runner scenario, we would we be incorrect in thinking that the murder of those replicant was morally wrong, given that replicants are, according to the social context of that fictitious world, just machines and are not appropriate subjects for moral consideration. Despite that, killing sentient, self-aware androids still, intuitively, strikes us as wrong. Clearly in environments where the social paradigm caters to injustice, this method is essentially useless as a method for delineating moral concern.

IV. Possible Objections and Responses

One issue often brought up when discussing moral status for machines is the concern that machines cannot feel pain or cannot be hurt or harmed. This is often coupled with doubts that machines can suffer, which of course is a major point of interest in the question of moral patiency.[41] The ability to suffer suggests a capability for states of wellness and distress within an entity, as well as basic needs for the entity that when taken away or provided either damage or bolster the well-being of said entity. It seems problematic that something that cannot suffer could be a moral patient given that moral patients are characterized by their needs, and by extension, their ability to suffer when deprived of those needs. It's important to recognize, however, that the capacity to experience pain is itself not broad enough to encompass our current moral community. Erica Neely uses an example of a person with congenital analgesia i.e., someone who cannot register pain to address this issue.[42] Even if we understand that this person cannot feel pain, moral intuition tells us that kicking this person is wrong, nonetheless. This is not because the action caused pain (because by definition, it did not). The proper way to understand why this is wrong is with respect to interests. What is wrong about kicking the man with congenital analgesia has to do with the damage that this kick could have created and presumably the affront to his need to remain unharmed. The attack not only causes harm, it inflicts harm that is unwarranted given the circumstances and is therefore morally impermissible. The results represent violations of the victim's interests and therefore, by definition, represent an injury to

[41] Neely, Erica L. "Machines and the Moral Community." p. 2.
[42] Neely, Erica L. "Machines and the Moral Community." p. 2-3

victim's well-being. By this model, so long as it is possible to harm the interests of an entity, it is also possible to harm the entity itself.

This falls into direct contradiction with Robert Sparrows' claim that a machine which does not suffer (in the sense that suffering still implicates states of pain and pleasure) "cannot be [an] appropriate object for moral concern at all."[43] Rather than insist that the capability to experience suffering does not play any part in moral consideration we could alter our meaning of suffering to accommodate our use of the term harm, such that suffering will also refer to the experience of having our interests thwarted. Additionally, we can be sure that a being lacks moral standing if it lacks interests. A crude example Neely uses in her work are chairs and tables.[44] They are objects with no clear needs, or states of satisfaction or dissatisfaction and so they don't have interests, which makes harming them, at least with respect to our use of the word, impossible.

Another question that will likely be raised after reviewing this argument is how my argument can claim that sentience is a necessary sufficient requirement of well-being but also claim that ultimately in judging whether a machine has sentience, genuine sentience no longer means anything and that if a machine is designed well enough to fool a moral agent they should treat the machine as a moral patient anyway.

To this I respond in agreement. The limitations we face in judging machines' capacity for sentience, self-awareness and essentially any non-apparent feature are considerable. We can never be sure that a machine's programming is the star of the show and in reality its mental

[43] Sparrow, Robert. "The Turing Triage Test."
[44] Neely, Erica L. "Machines and the Moral Community."

facility amounts to little more than a calculator. We also have no way of confirming that any phenomenal experiences within a machine are legitimate. All we have are these apparent features displayed by our machines and our impression of what it means to have moral status. It seems best to recognize our limits and adopt a method of examination that is effective as possible given what we have to work with.

Conclusion

This account of moral status recognizes the incredible endeavor we face in attempting to uncover genuine representatives of sentience among machines. Rather than build a method which ignores the pitfalls of moral status in machines, this project was designed to capitalize on them. For example, rather than assert that sentience is apparent objectively even in entities with behaviors as dubious as machines, this argument found support in the concept of apparent and non-apparent features which played well with the idea that machines could conceivable behave so convincingly like a moral patient and yet not adhere to any of the typical internal behaviors defined by sentience.

V. References

1. Sparrow, Robert. "The Turing Triage Test." Ethics and Information Technology, vol. 6, no. 4, 2004, pp. 203–213.

2. Block, Ned. "Introduction: What Is Functionalism?" The Language and Thought Series, 2006.

3. Anderson, David Leech. "Introduction to Functionalism." 2003.

4. Coeckelbergh, Mark. "Robot rights? Towards a social-Relational justification of moral consideration." *Ethics and Information Technology*, vol. 12, no. 3, 2010, pp. 209–221., doi:10.1007/s10676-010-9235-5.

5. Neely, Erica L. "Machines and the Moral Community." *Philosophy & Technology*, vol. 27, no. 1, 2013, pp. 97–111.

6. Sparrow, R. 2012. Can machines be people? Reflections on the Turing triage test, in *Robot Ethics: The Ethical and Social Implications of Robotics (*ed. P. Lin, K. Abney, and G.Bekey), MIT Press, Cambridge, Mass., 301-315.

7. Selyukh, Alina. "Tech Giants Team Up To Tackle The Ethics Of Artificial Intelligence." NPR, NPR, 28 Sept. 2016,

8. The Future of Humanity: Heirdegger, Personhood and Technology, Mahon O'Brien

9. Legal Personhood for Artificial Intelligences, Lawrence B. Solum

10. Harry G., Frankfurt, "Freedom of Will and the Concept of a Person." *www.jstor.org*. Journal of Philosophy, Inc., Vol. 68, No 1.

11. Sparrow, R. 2012. Can machines be people? Reflections on the Turing Triage Test. In Patrick Lin, Keith Abney, and George Bekey (eds) Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, Mass.: MIT Press, 301-315.

12. Warren, Mary Anne. "Moral Status: Obligations to Persons and Other Living Things." Oxford Univ. Press, 2009.

13. Cochrane, Alasdair David Charles. "Moral obligations to non-Humans." University of London.

14. Scott, Riddley, director. Blade Runner . Warner Brothers, 1982.

15. Feinberg, Joel, 'The Rights of Animals and Unborn Generations" in William T . Blackstone (ed.), Philosophy and Environmental Crisis, (Athens: The University of Georgia Press, 1974).

16. Wilkinson, Stephen, "Bodies for Sale: Ethics and Exploitation in the Human Body Trade", (London:Routledge, 2003).

17. "Criteria for Recognizing Sentience." Animal-Ethics.org

18. Turing, A. M. "I.—Computing Machinery And Intelligence." Mind, LIX, no. 236, 1950, pp. 433–460.

19. Nussbaum, Martha C., "Objectification", Philosophy and Public Affairs, Vol. 24, No. 4, Autumn 1995: