**An Account of Moral Standing for Machines**

Randy Bruno-Piverger

Florida State University

The question of what constitutes a moral being is an important one, particularly for those involved in the development of artificial intelligence. Researchers and industry leaders in the field of AI understand the awesome potential this technology holds. In recognition of this potential many of these industry leaders have come together, onto ethical boards and associations, with the goal of confronting the dangers of AI development.[[1]](#footnote-1) Of the many questions groups such as these will encounter there is one that many at the crossroads of ethics and computer science find deeply troubling, i.e., at what point can we say that we should have moral consideration for a machine?

Robert Sparrow outlines relevant factors for future judgements on the intelligence of a machine as well as a test to determine when a machine has achieved moral standing comparable to that of a human. He argues that passing Alan Turing’s “imitation game” is more than sufficient to prove a machine intelligent; the test also demonstrates a machine is self-conscious, and has the ability to form projects and hold ambitions.[[2]](#footnote-2) 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. His test is a moral dilemma. It is a situation 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.[[3]](#footnote-3) 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.[[4]](#footnote-4) 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.[[5]](#footnote-5) Moreover, Sparrow asserts that we can know whether a machine’s moral standing is comparable to that of a human by conducting his Triage.

In this work, I argue that the essential requirement for moral standing should rest on the capability for a machine to have interests, and not necessarily the ability to experience pain or pleasure. Moreover, I argue that Sparrow’s test is a good start with respect to how we ought to be thinking about moral consideration for machines but ultimately moral comparisons made on the basis of a personal experiences, like in his Triage, are best explained in terms of a social-relational experience model. In Section I, I will introduce an argument from Erica Neely that prioritizes the capability for having interests as a requirement for moral standing. I will also explain why it is preferable to the requirements referred to by Sparrow, i.e. the capability to sense pain or pleasure. Then in Section II, I will introduce a basic framework for social-relational moral justification from Mark Coeckelbergh and explain why, in this case, it is preferable to view our moral relationships through such a framework. Section III will review likely rebuttal arguments.

I.

Erica Neely claims that what is needed to have basic moral standing can be understood in terms of whether or not a being has interests. Specifically, Neely uses a definition by John Basl that describes interests as things that when satisfied “contribute to [an individual’s] welfare.”[[6]](#footnote-6) By understanding interests in this way we are able to approach some of the most troublesome issues involving machines and moral consideration. One such issue is the concern that machines cannot feel pain or cannot be “hurt.” 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.[[7]](#footnote-7) It’s important to recognize, however, that the capacity to experience pain is itself not broad enough to encompass our current moral community. Neely uses an example of a person with congenital analgesia i.e., someone who cannot register pain.[[8]](#footnote-8) 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). A better way to understand why this is wrong is with respect to interests. What is actually wrong about kicking the man with congenital analgesia has to do the damage that this kick could have created, and presumably the disruption of his desire to remain unmolested. Both of these results represent violations of the victim’s interests. 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 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.”[[9]](#footnote-9) Rather than insist that the capability to experience suffering does not plays any part in moral consideration we could alter our meaning of suffering to accommodate our use of the term harm. In this way suffering can also refer to the experience of having our interests trod upon. Generally we can associate interests with entities that have goals, desires, needs, or states of satisfaction or dissatisfaction.[[10]](#footnote-10) These are also not collectively strict requirements. It is enough for an entity to have goals, and not necessarily have sentience, or vice versa, for that entity to have interests. 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.[[11]](#footnote-11) 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.

II.

Now that we have a minimal necessary requirement for moral consideration, we might continue our inquiry by asking how do determine the degree to which we apply moral consideration. In a similar forum of discussion, Mark Coeckelbergh argues that there is another, widely unconsidered, approach for moral consideration of robots, animals, and humans that can help us answer this question. This method presents a “social-relational” justification of moral consideration.[[12]](#footnote-12) 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 structure.[[13]](#footnote-13) 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.[[14]](#footnote-14) 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.”[[15]](#footnote-15) Third, the experiences 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.[[16]](#footnote-16) The recognition that our observations of the world are to some extent internalizations and that there is no observer-independent reality or “thing-in-itself,” also motivates this course thinking. This 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).[[17]](#footnote-17) Together the subject-object dependency implies that moral significance arises from the relation between the object and the subject. Lastly, 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”[[18]](#footnote-18) the moral arguments we engage in, but also those interactions emerge subsequent to some social context. 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 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.[[19]](#footnote-19)

When trying to determine what degree of moral consideration is appropriate for a particular entity, the social-relational method asks that we consider the apparent features we notice in an entity, then consider then current social context that we are a part of. We then compare 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:

To understand how one would manage their moral considerations to their dog, who for the sake of this example suffers from separation anxiety, they 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 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 to 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 apparently 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 your dog clearly does not enjoy being left alone it seems reasonable to confer some moral consideration to your animal; Perhaps by adopting 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 is not radically different from Sparrows own Triage where he 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 bestowing moral consideration to a machine—considerations comparable to that of a human. This 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 major weakness this method also 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. Biased or “idiosyncratic” individuals or those with less than average reasoning ability should not be considered proper subjects for the test.[[20]](#footnote-20) 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. Our social-relational method revokes this idea as moral consideration exist 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 our ideas on apparent features and the internalization of our observations. With those ideas in mind, 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.

III.

The method I have outlined here is not without its weaknesses. One question that should be answered immediately is how the minimum necessary requirement for moral sanding, the capability to have interests, is related to the tenets of the social-relational method. Moreover, is the observation of interests in another entity, by our own definition, just another apparent feature of some entity and by extension just another subjective observation? The answer is yes, this method works by recognizing the personal subjectivity and social influence involved in moral consideration. The hope is that this model can help depict how moral consideration already functions not how it ought to function. With that established it is important to also note that the variations in moral values across time and cultural settings is best explained by the influence of social paradigms and the nuance of individual subjective impressions on morality.

To return to the question of how this model is related to interests, it is also true that by this model any mention of observing interests or the capacity of interests in an entity amounts to another apparent feature. By our own social context, at least what I gather from it, the observation of interests (or other capacities that indicate an entity has interests), always precedes moral consideration, making it at least worth mentioning here as a possible prerequisite for the human behavior of moral consideration. I have yet to find a case that convincingly demonstrates anything to the contrary. I cannot, however, say that this will be the case for everyone. Moral consideration will continue to play out on a person-to-person basis and, in virtue of its connection to social context, will continue to evolve.

IV.

The social-relational model for moral consideration provides a demanding but flexible method to determining moral consideration. As I have argued, it is superior to the Triage Sparrow proposes thanks to its provisions for assigning varying degrees of consideration, and it’s recognition of the roles of social influence and personal subjectivity in moral consideration. Apart from this, concerns with the threshold between moral consideration and the lack thereof (“at what point should a machine have moral standing?”) have been addressed through the introduction of interests.

References

1. Sparrow, Robert. “The Turing Triage Test.” Ethics and Information Technology, vol. 6, no. 4, 2004, pp. 203–213., doi:10.1007/s10676-004-6491-2.
2. Block, Ned. “Introduction: What Is Functionalism?” The Language and Thought Series, 2006, doi:10.4159/harvard.9780674594623.c17.
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., doi:10.1007/s13347-013-0114-y.
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. <http://www.jstor.org/stable/2024717>
11. 55 Emory L.J. 347 (2006) Chimera and the Continuum of Humanity: Erasing the Line of Constitutional Personhood, Bennett, D. Scott [42 pages, 347 to 388 ]
12. *Philippe Caillou, Benoit Gaudou, Arnaud Grignard, Chi Quang Truong, Patrick Taillandier.*

*A Simple-to-use BDI architecture for Agent-based Modeling and Simulation. The Eleventh*

*Conference of the European Social Simulation Association (ESSA 2015), Sep 2015, Groningen, Netherlands. <http://www.essa2015.org/>. <hal-01216165>*

1. 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.

1. Selyukh, Alina. “Tech Giants Team Up To Tackle The Ethics Of Artificial Intelligence.” [↑](#footnote-ref-1)
2. Sparrow, Robert. “The Turing Triage Test.” [↑](#footnote-ref-2)
3. Sparrow, Robert. “The Turing Triage Test.” [↑](#footnote-ref-3)
4. Sparrow, Robert. “The Turing Triage Test.” [↑](#footnote-ref-4)
5. Sparrow, Robert. “The Turing Triage Test.” [↑](#footnote-ref-5)
6. Neely, Erica L. “Machines and the Moral Community.” [↑](#footnote-ref-6)
7. Neely, Erica L. “Machines and the Moral Community.” [↑](#footnote-ref-7)
8. Neely, Erica L. “Machines and the Moral Community.” [↑](#footnote-ref-8)
9. Sparrow, Robert. “The Turing Triage Test.” [↑](#footnote-ref-9)
10. Neely, Erica L. “Machines and the Moral Community.” [↑](#footnote-ref-10)
11. Neely, Erica L. “Machines and the Moral Community.” [↑](#footnote-ref-11)
12. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-12)
13. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-13)
14. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-14)
15. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-15)
16. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-16)
17. ­­­ Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-17)
18. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-18)
19. Coeckelbergh, Mark. “Robot rights?…” [↑](#footnote-ref-19)
20. Sparrow, Robert. “The Turing Triage Test.” [↑](#footnote-ref-20)