**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 a superficial grounds for moral concern on the basis of 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, on its ability to have personal projects and on its rational capabilities. 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,

1. 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.

2. 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 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.

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.” 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.[[5]](#footnote-5) 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. 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.” 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.[[6]](#footnote-6) These are also not collectively strict requirements. It is enough for an entity 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. 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. 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.[[7]](#footnote-7) 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. 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,”[[8]](#footnote-8) that is to say, “features-as-experienced-by-us.” 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.[[9]](#footnote-9) 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 (Locke).[[10]](#footnote-10) 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”[[11]](#footnote-11) 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.[[12]](#footnote-12)

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 know, (at least apparently) that a dog can experience 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 to be invested in the health of our animal, 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 and fair 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 stating 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. This is problematic for a number of reasons. For one, the idea of implementing quality control on moral consideration is troubling 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 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, if any moral consideration is ultimately a 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, then 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.

II.

Using the social-relational model in a similar fashion to Sparrow’s Triage can give us and the relevant qualities laid out by Sparrow we can at least begin to form a description of what such a being might require to be considered of moral worth. For one, this machine would require a self-awareness. We might describe this functionally as a mechanical state in which our machine is able to recognize itself as distinct from its environment, is capable of expressing self-interest and capable of spontaneously forming projects based on that expression, and finally that this state persists through the life of the machine. We might also require that this machine must have some states for pain and suffering as well as pleasure and contentment. We count develop a machine systems which mirror our own body’s organs to create this. For example, to reproduce neurological pain we install a synthetic nervous system which behaves like our own and initiates states of painful sensations or pleasing sensations according to our own highly detailed functional requirements. As for intelligence, we need persistent causal functions engaged in observation and general learning. Such a state would collect general and specific observations about the world and when connected to our functional state of self-interest would be able to causally trigger a state such that abstract design and strategy could be conceived and implemented.

The brilliance of complying with this model is that it we can continue to add functional details to a point of precision that will satisfy most people, even if at the end of the day this machine’s hardware and internal functions are physically very different from our own, their functional roles will be nearly the same.

III.

Functionalism, of course, is not without its weaknesses. Most quickly some will realize its most glaring issue is an attack via reductio. This spoils functionalism here in a few ways. First, no matter how close we can claim our functional description is to the truth there may always be an addition to make, a detail missed, a causal role overlooked. For example:

Pain, as we described earlier, might be simply described as a causal event between the firing of certain nerves resulting in an unpleasant neurological sensation or mental state. But alas, pain should also be memorable, to help avoid the cause of that pain in the future. Thus any pain state must have a causal relationship with our memory faculties. But hold on, pain should also, in some cases move us to tears, and it also needs a causal relation to our tear ducts—and so on.

We can keep building indefinitely and we may always end up lacking completeness. I cannot doubt the power of such a reductio. In my defense, however, the observable configuration of a human, her physiology and chemical make-up are finite. Which should mean that if this mental programming is ever perfected it must also have finite qualities.

The next argument from reduction is more troubling than the last. It suggests that no matter how specific we are with our functional requirements, imagining a machine who could imitate the outward expression of those states without ever experiencing them will always be possible.[[13]](#footnote-13) This is conceivably true, however, functionalism’s defense against this is simply its openness to being as descriptive as we can about what we expect from our systems functionally. The more detail we issue a definition the more difficult it is to create an exception. As stated previously, it allows us to create functional definitions with precision unbound; only limited by the content of our own observations.

IV.

Functionalism revolutionized contemporary theory of the mind by giving us an account of mental properties in terms of their function as opposed to their composition. This opened the door for empirical researchers of the mind as well as philosophers looking to form an account of the mind less reliant on human physiology. I have argues that Functionalism can provide a sturdy functional definition of not only intelligence within a machine but also a functional definition of moral worth within a machine.

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. Selyukh, Alina. “Tech Giants Team Up To Tackle The Ethics Of Artificial Intelligence.” NPR, NPR, 28 Sept. 2016,
5. The Future of Humanity: Heirdegger, Personhood and Technology, Mahon O’Brien
6. Legal Personhood for Artificial Intelligences,  Lawrence B. Solum
7. 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>
8. 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 ]
9. *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.” 2016 [↑](#footnote-ref-1)
2. Sparrow, Robert. “The Turing Triage Test.” 2004 [↑](#footnote-ref-2)
3. Sparrow, Robert. “The Turing Triage Test.” 2004 [↑](#footnote-ref-3)
4. [↑](#footnote-ref-4)
5. [↑](#footnote-ref-5)
6. [↑](#footnote-ref-6)
7. [↑](#footnote-ref-7)
8. [↑](#footnote-ref-8)
9. [↑](#footnote-ref-9)
10. ­­­ [↑](#footnote-ref-10)
11. [↑](#footnote-ref-11)
12. [↑](#footnote-ref-12)
13. Block, Ned. “Introduction: What Is Functionalism?” 2006 [↑](#footnote-ref-13)