**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 a 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. According to Sparrow 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. 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, 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 comparisons made on the basis of moral standing, between humans and machines (e.g. Sparrow’s Triage), are best explained in terms of a social-relational experience.

To answer such a question we must first recognize an implicit assumption the question relies on, that is, moral consideration is something applicable to non-human entities. This is an idea that Erica L. Neely defends in her work “Machines and the Moral Community.” She claims that what is needed to have basic moral standing can be understood in terms of whether or not a being has interests. She endorses a definition by John Basl that describes interests as things that when satisfied “contribute to [an individual’s] welfare.” Neely argues that conscious machines must have moral standing. She defines conscious as a state where one is capable of experiencing things in either a positively or negatively valenced way and be able to have desires, needs, goals, and states of satisfaction and dissatisfaction. Since Nelly’s conscious beings have needs and goals that contribute to their wellbeing, it is clear that they have interests, and also moral standing. Generally, these qualities represent the requirements Neely uses to determine whether or not a particular being has interests. Neely expands on what it means to do harm against another being with respect to interests. A common worry when considering how we might include machines in our moral considerations is that machines that cannot feel pain cannot be “hurt.” This is often coupled with doubts that machines can suffer, which is a major point of interest in the question of moral patiency.[[4]](#footnote-4) Neely describes this characterization of harm as an internalization of sentience. Moreover, it is also not broad enough to encompass our current moral community. Neely uses an example of a person with congenital analgesia, a person who cannot register pain. Even given the fact that this person cannot feel pain, it seems obvious, that kicking this person would still be wrong. Neely points out that this is not because the action caused pain (because it did not). What is wrong about kicking this person has to do the damage that this kick could have created, and the disruption of this person’s desire to remain unmolested. Both of these results represent violations of the victim’s interests. Additionally, we can be sure that a being lacks moral standing if it lacks interests. An example she uses often in work text are chairs and tables. Clearly they don’t have interests, which makes harming them, at least with respect to our use of the word, impossible.

In a similar forum of discussion, Mark Coeckelbergh argues that there is another, widely unconsidered, approach for moral consideration of robots, animals, and humans. 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.[[5]](#footnote-5) 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,”[[6]](#footnote-6) 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.[[7]](#footnote-7) This is a direct 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).[[8]](#footnote-8) 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 interactions between subject and object exist “prior to”[[9]](#footnote-9) the moral arguments referenced by Coeckelbergh[[10]](#footnote-10), but also emerge subsequent to some social context. Simply put, the social context which greatly affects our moral considerations 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.[[11]](#footnote-11)

I.

There are two crucial aspects of “functional analysis” I will introduce here.[[12]](#footnote-12) They are the specificity of the causal relationship you are describing, and the uniqueness of the mental state this interaction inspires. To develop a useful method of functional reasoning we might start simply by looking closely at what it is we are trying to describe. This will typically involve asking the question: “What functional quality or qualities do things of this sort have in common?” or “What makes them distinct?” It is important to be sufficiently specific when describing an object or risk an overly applicable definition. Take an orange, for example:

Functionally we will say that an orange is object that once consumed provides various us with various nutrients, these nutrients (when consumed in large enough quantities) influence in us a mental state of satiation. We can eat oranges and feel full afterward. This much is true of oranges but it is also true of apples, or dates or mangos.

We have made the mistake of being too broad with our functional definition. Satiation may still be part of a reasonable definition for this fruit but it is obvious we need more information. A better example:

An orange is object that once consumed provides various us with various nutrients, these nutrients (when consumed in large enough quantities) influence in us a mental state of satiation. Moreover, when eaten the orange provides a very specific neurological response, triggering immediately a sensory experience of sweet or sour acidic flavor. Visually it is capable of providing our senses with a specific range of color experiences (green, yellow, and orange colors) not to mention it spherical shape which will also engender a unique mental state in us upon observation.

This collection of functional describers provide a more detailed and more meaningful idea of what it means to see and eat an orange, and by functionalism it serves as a more appropriate definition of what an orange is. This is a good example because it is strict enough to avoid having other objects with no business being compared to an orange, matching our description. It is also a good example because it allows for the orange to be multiply realized, that is to say, on functionalism, an orange can be realized in a number of ways.[[13]](#footnote-13) This is a major strength of this view as it escapes the species bias physicalists must deal with. Just the same, there are a number of fruits that vary in flavor, color and size that we would also call an orange. There is the common orange, the navel orange, the seville orange, and many more. Indeed, there is often a range of acceptable variation in our definitions of things. Functionalism addresses the need for this variation quite nicely. A general rule we can from these examples is that finding specific functional qualities that are common to all verities of an object you are describing is an essential part of using this method successfully. To highlight again how important this rule is we can examine Turing’s and Sparrow’s very broad functional requirement set forth in his Triage Test.

Turing’s Test would have us believe that is a machine can fool us into thinking they are human by successfully impersonating a human in a conversation, that we should consider this machine intelligent. Functionally Turing’s Test requires a machine to interact with a human such and causally influence states within the human that resemble perfectly such states that would appear if said human were speaking to any other human. Sparrow also insists that passing the test would also mean that this machine is self-aware and capable of forming projects and ambitions. There is, however, a stifling problem here. Is the ability to report on *some* internal state or *some* ambition or project (but not necessarily your own), be sufficient for us to suspect that a machine is intelligent, more remarkably is it enough for us to consider said machine self-aware? One thing is certain clear, by Sparrow’s own account, a machine could be developed that can pass Turing’s and his own Triage Test without necessarily having the faculties those tests aim to reveal. It is not hard to envision such a machine. A learning machine with enough data and enough time to practice could conceivably get a very good handle on the human language. If our machine had access to, for example, a database containing every conversation a man took part in over the course of his life, was tooled with an advanced learning algorithm, and was set aside for an indefinite period of time to train, whereby the machine is continuously given feedback on the quality of its responses and its ability to continue a conversation convincingly over time—it would only be a matter of time until this machine completes the intended course of its programming. Another example:

Imagine a machine capable of building a profile for its conversations based entirely off information it has collected from previous conversations. This profile would act as an immediate reference for questions like: “how are you doing today?” or “what are your plans for the year?” The machine would be capable of answering questions inquiring as to its internal states without any kind of observation on its true internal states ever occurring. Its projects would be simply an output value, fished out from millions of other potential responses.

There is no reason for us to think that such a machine could not exist. This example again demonstrates how designing functional requirements too broadly can ruin a functional definition altogether.

As for the requirements set forth by Sparrow’s Triage, they too could benefit not only from greater specificity but also from recognizing the necessity of unique causal roles. In his Triage, Sparrow requires that a machine interact with a human in a situation of moral difficulty such that it causally influences states within the human that resemble perfectly such states that would appear if said human were interacting with any other human. This test is stronger than Turing’s in at least one way: that it is seemingly more difficult to convince a human that allowing a machine to “die” is every bit as morally damnable as letting a human die than it is to convince a human that they are speaking to another human.[[14]](#footnote-14) This difficulty puts a heavier burden on what the machine must be capable of functionally performing. The test’s downfall, however, is its reliance on human sentiment. Much like Turing’s test, this Triage is making a claim that human sentiment can set some kind of standard for moral worth (in Turing’s case it was intelligence). This idea of course is very much at odds with the spirit of functionalist analysis—that is the capability of multiple realization and the avoidance of species bias. It may be the case that humans are not capable of extending equitable moral standing. It is easy to conceive of a situation where a human embraces there species bias and will not be overly burdened by the choice of a letting a machine die. Simply knowing that the machine is not human would be enough. Human already behave in this way; not with machines but with other humans, along lines far less controversial than species (race, economic status, familial relation, etc.). Even in the presence of a morally worthy machine, this test fails because humans in general cannot be relied upon to consistently make correct moral decisions. More worry some than this is the human who is fooled by a machine with no real claim to moral worth. Human bias is exploitable, and designing a robot to look exactly like a human speaks to the depths of this bias. This concern is led by the same issue that Turing’s test faces. What of machines that are designed specifically to illicit moral concern from us but in reality have no intelligence, no sense of self and no understanding of morality? Though they would be harder to design their prototypes already exists, in video game simulations. These simulations are designed to and succeed in evoking emotional attachment to “weak AI” avatars.[[15]](#footnote-15) Their narratives tug on our sympathies and motivate us to invest emotionally in these avatars. All the while, we realize that they are not human nor intelligent or aware. The human proclivity to sympathize liberally and the opposite should worry us, if our aim is to design a test for moral worthiness in other species. We cannot expect that what is unique about morally worthy beings is the reaction that moral worth triggers from those outside the being. Particularly if those reactions are of an inconsistent nature. For the sake of the test we must find causal roles and ultimately mental states that function with an algorithmic consistency.

For this reason, if machines are to achieve any kind of moral standing it needs to arise from a reasonable recognition of the functional capacities within them and not from their ability to fool or otherwise enchant their human observers. There are a few more reasons to agree with this. By exploring what is required of any person to be of moral worth we learn more about ourselves as a species and ideally we develop the ability to recognize it as a trait in other species. It should not matter that we are convinced via a machine’s human like performance. What is more important is the discovery of these causal roles that influence the states we cherish most in our human experience; those involved in moral duty, empathy, love and friendship, as well as many other.

II.

Using my model 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.[[16]](#footnote-16) 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. Block, Ned. “Introduction: What Is Functionalism?” 2006 [↑](#footnote-ref-12)
13. Anderson, David Leech. “Introduction to Functionalism.” 2003. [↑](#footnote-ref-13)
14. Sparrow, Robert. “The Turing Triage Test.” 2004 [↑](#footnote-ref-14)
15. Sparrow, Robert. “The Turing Triage Test.” 2004 [↑](#footnote-ref-15)
16. Block, Ned. “Introduction: What Is Functionalism?” 2006 [↑](#footnote-ref-16)