Thomas O’Callaghan-Brown

Why the contemporary view of the relationship between moral status and rights is wrong, with specific reference to the treatment of AI

The contemporary view of moral status and rights is that for an entity to qualify for rights that being must have moral status. There are many different criteria for what constitutes a being having moral status, but the general framework, when applied to humans is as follows:

P1[1]: Iff (if and only if) an entity has moral status does it qualify for moral rights

P2: Humans of all forms have moral status

C: Humans of all forms qualify for moral rights (□ Elim P1, P2)

When we apply this to AI there is some uncertainty, but authors such as John P. Sullins and Colin Allen suggest that AI of all kinds have moral status. Following from this we can assume that they are owed moral rights. As such the conditional stated above, with reference to AI would be posed as:

P1: Iff (if and only if) an entity has moral status does it qualify for moral rights

P2: AI of all forms has moral status

C: AI of all forms qualifies for moral rights (□ Elim P1, P2)

I would like to suggest that it is indeed true that AI deserve moral rights, but that they cannot be said to have moral status. Therefore, a change to the framework ought to be carried out. This change is necessary due to the confusion that arises out of labelling AI and other

[1] This is a form of laying out an argument whereby P stands for a premise or step in the reasoning, and C is the conclusion that follows from it.

entities as moral agents, which can be avoided by arguing that no entity needs moral status to be treated morally. I will demonstrate this by systematically showing that P1 and P2 are incorrect using Susan Leigh Anderson's Kantian approach regarding indirect moral duties to refute the necessity of moral status; and Searle's Chinese Room with Fodor's physicalist critique to refute the moral status of AI. Following this I will amalgamate the two views to show why AI, despite their non-moral status, ought to be owed moral rights. Finally, I will suggest emendations to the contemporary framework and the consequences this has for other non-moral agents.

Before I evaluate the conditional statement, I think it is important provide reference to a specific author who espouses this view of moral rights. It is easy to simply say that this is what people think, but to do so is spurious and lazy. By providing a specific author it gives this view a concrete grounding and directs the argument more. In his 1991 article, Tibor R. Machan, an ethicist who Anderson takes aim at in her writing, suggests that animals do not have rights owing to their lack of moral agency. He writes "*there is plainly no valid intellectual place for rights in the non-human world, the world in which moral responsibility is for all practical purposes absent"* (Machan, 1991). Clearly Machan maintains a view that for a being to have rights it must have a moral status. His view is that status is only invoked by the ability to act morally and can be referred to as a "Moral Agency" account of moral status. As we shall see this is one of many views of moral status and its prerequisites. In any author that suggests a theory of moral status there is an implicit (or even explicit) commitment to the framework indicated.

The issue with the first premise can be framed by the complications that arise concerning status. In her book concerning the topic, Mary Anne Warren divides the views of moral status by what she calls "*intrinsic properties [...] identified by some philosopher as the single necessary and sufficient condition for the possession of moral status"* (Warren, 1997). These are identified by Susan Leigh Anderson as "*Sentience, Emotionality, Reason, the capacity to communicate, Self-awareness, and Moral agency"* (Anderson, 2011). Already we can see that this will be problematic – philosophers have been, and still are, attempting to find a specific aspect on which they can place status that is coherent and produces valid and desirable outcomes. We can take a look at various accounts of moral status and see which hold and which don't.

The Sentience Account states that a being deserves moral status in so far as it can feel pleasure and pain. This is proposed by Utilitarian's in general, Jeremy Bentham, the so-called Father of Utilitarianism, wrote "*the question is not, can they reason? nor, can they talk? but, can they suffer?"* (Bentham, 1823), this view was also taken on by the Preference Utilitarian Peter Singer who argued for "*the equal consideration of interests"* (Singer, 1995). Singer argued that to focus on humanity alone in terms of moral consideration is "speciesist*"*. This term was coined by Richard Ryder who discussed it in his 1970 leaflet on the subject. He said that "*it is illogical to argue that other animals do not suffer in a similar way to ourselves"* (Ryder, 2010), and if we do then we fall victim to speciesism – that is prioritising our own species over others. This is a problem though, with reference to Artificial Intelligence, as we cannot argue (at least at the moment) that they experience pain, while they may be able to eventually, the logical conclusion of the Sentience account is the amoral status of AI. Let us consider the genetic abnormality: congenital insensitivity to pain. While this condition has only afflicted 20 people according to scientific literature (Dabby, 2012), and this is an exceptional case, according to the Sentience account, their inability to process pain would render these individuals void of moral status. This seemingly provides a counterexample to the Sentience Account of moral status, where humans are rendered void of moral status based on a deficiency out of their control.

Another view of moral status of is that of Self-Consciousness. This is often attributed to Kant, as it is through a sense of self-consciousness that we think of ourselves as the authors of our own thoughts, beliefs and actions. And as a result, we can be viewed as autonomous – we are not aware of any being other than ourselves imposing on our actions and in being free we are able to act, and therefore claim moral status to Kant. But even this is not flawless. First, what ought we to say of infants that are too young to possess such faculties? Are they now void of moral status until they come of an age to be considered self-conscious? Further, there are animals other than us that we know are self-conscious through a task known as the Mirror Test, developed by Gordon Gallup Jr in 1970. These include ravens, dolphins, great apes and several others. Ought we then to include them under the umbrella of moral status but not other animals? Once again, we are at a disjunction and the theory seems to fall apart.

The final view I shall evaluate is that of autonomy. This is based off of the Kantian idea in a way as well, but I think that this is a more valid argument. Kant founds his view of morality in the view that for an action to be moral it mustn't be at all heteronomous - or imposed upon by some other force. Andrew Reath, a commentator of Kant, described his commitment to the autonomy of the will as "*the independence of the rational will from externally imposed principles and its capacity to generate authoritative norms"* (Reath, 2006). But Kant is committed to this due to his ontological beliefs. If we look at the idea of autonomy psychologically it has important implications for morality too. In being free we are therefore culpable or worthy of praise. If you are compelled into performing an action you are not viewed as being culpable, take for instance someone driving a car that has, unbeknownst to them, had the brakes cut. They are heteronomously bound by the limits of the car, in other words no matter how free they are the car imposes its inability to brake on them. So, if they tried to stop and ended up hitting and killing someone we would not say it was their fault as the car impinged upon their freedom to stop and it is the *car* that is to blame. In this way we can see that the autonomy of the agent plays a big role in assigning praise or blame. In terms of morality, we cannot say that an action is good or bad (i.e. morally praiseworthy or morally blameworthy) unless the agent was free to choose to do that action. Animals cannot be viewed as free to choose their actions as this requires a cognitive faculty of projection: being able to imagine what could happen. This itself has an implicit idea of self-consciousness as necessary – it is the ability to not only be aware of yourself but to be able to put yourself into situations that could happen. Thomas Nagel spoke about this as an essential capacity of consciousness, the "*subjective character of experience"* or a "*what is it like to be X"* experience (Nagel, 1974). This imaginative act of projection is how we determine whether we wish to do an action or not. So, it seems to me that this is the most feasible explanation. It seems to align with the methods by which we choose to act in our day to day lives and doesn't limit itself to one facet of our decision making.

Even with this theory there is the problem of exceptions. Like the previous view of moral status, it doesn't include children, and other people incapable of expressing autonomy – coma patients for instance. This is where I would like to bring in Kant's view of moral duties with respect to entities, a view mentioned in his *"Lecture on Ethics"* (Kant, "Our Duties to Animals" in Lectures on Ethics (Infield, L. trans. 1963), 1780) and built upon by Susan Leigh Anderson in her paper "*The Unacceptability of Asimov's Three Laws of*

Robotics" (Anderson, 2011). Seeing as the importance of Autonomy and its role in morality is a commitment of Kant, it only makes sense that we continue in his line of thinking. Kant maintained that the treatment of animals is indicative of how we treat humans. He saw animals as deserving of moral rights with respect to them. These are indirect moral duties as opposed to direct moral duties – the latter being those owed to other human.

On the topic of animals Kant wrote "*animals [...] are there merely as a means to an end. That end is man"* (Kant, "Our Duties to Animals" in Lectures on Ethics (Infield, L. trans. 1963), 1780). This may seem to be portraying animals as subhuman and therefore contradictory to the view that animals are owed indirect moral rights, but I think it would now be pertinent to look at his idea of the categorical imperative with regards to morality to show this is not the case.

The categorical imperative is a tool Kant developed to ascertain the morality of any given action. It is prescriptive regardless of your ends and ought to be followed no matter what. Here we see Kant's commitment to duty. This view is a deontological one, and can be summarised as a theory that *"guide*[s] *and assess*[es] *our choices of what we ought to do, in contrast to those that guide and assess what kind of person we are and should be"* (Alexander & Moore, 2016). Kant argued that "*wherever possible we have a duty to promote the highest good"* (Kant, 1998). It is this sense of duty – performing the right action for the right reason regardless of your own subjective ends – that Kant grounds morality in. Returning to the Categorical Imperative, Kant's duty extends to this too: using the categorical imperative to determine whether an action is right or wrong, and acting in accordance. The most relevant facet for this moral-duties-with-respect-to-an-entity idea is the second categorical imperative – the formula for ends in themselves – regarding the treatment of humans. He wrote "*Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end"* (Kant, 1785). Now if we take a look at the way Kant described animals earlier: "*as a means to an end"* (Kant, 1780), we can see that we cannot treat animals as a mere means. Their mistreatment is a violation of the categorical imperative and thus an immoral action.

Even more significant for Kant is the idea that animals ought to be treated well owing to their similarity to humans. He writes of dogs "*If a dog has served his master long and faithfully, his service, on the analogy of human service, deserves reward."* (Kant, 1780), such an action is indicative of how we treat other men. While animals are not, to Kant, moral agents nor are they beings with moral status, their shared traits with humanity warrant a just treatment not for their own sake but with respect to them and with respect to what they represent. In a later section I will show how this is just as applicable to AI. And so, we now see that it is possible, contrary to *P1* for an entity, lacking moral status by our definition, to be owed moral rights with respect to appearance and the way they act.

Next it is necessary to address the idea of *P2:* that AI have moral status. So far, all moral status accounts have dealt only with humans and animals, AI fall in a category somewhere in between. Unlike animals there are those who would argue that AI possess rational faculties and are capable of human-like cognitive processes; and unlike humans AI are created, they are programmed by a human executor, furthermore, there emerges a difficulty in categorising AI as self-conscious beings, a sort of grey-area. So to address whether it is the case that AI can be said to have moral standing I will address whether they can be considered rational and autonomous, as humans are.

If we look at the Bicentennial Man, a short story written by Isaac Asimov from the perspective of a self-conscious rational automaton, then it is clear that when an AI reaches such a level of cognitive abilities that it and a human are virtually indiscernible, we have no choice but to label it as human. The issue is, we can never know this for certain. The Turing Test described by Alan Turing in his 1950 paper *Computing Machinery and Intelligence* has the aim of ascertaining in terms of "*... one particular digital computer C.* [whether it is] *true that [...] C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?"* (Turing, 1950), where A and B are human participants. To this day this is the test for machine consciousness: whether it can be considered indiscernible from a human subject. Issues are often raised with this definition, arguing that a machine could be composed to pass the test without being actually self-conscious. The argument against it essentially boils down to the subjectivity of experience: as we cannot know the thought process of the machine, we can never be sure of its self-consciousness. This rejection of the Turing Test is significant especially when it comes to determining whether AI have moral status.

But Asimov, in his short story, responds to this objection with the character of Little Miss who says "*it is difficult to determine whether even another human being has feeling like oneself. All we can use are behavioural cues."* (Asimov, 1976). Here Asimov raises an interesting response to the objections of the Turing Test. The same critique that can be applied to AI by the objectors can be applied to other humans. We have no way of knowing whether the people around us think and reason the same way we do, we can only infer from behaviour. And so, we must do the same with AI.

There is a difference though between performing such an inference on humans and AI, as members of the same species it is easier to assume that those around us, having been created in similar ways and being made up of similar stuff, might think in a similar way. But with AI there is a reduction to the purely material makings of humanity, so to make an extrapolation of self-consciousness is far more unfounded. John Searle also disagreed with the validity of the Turing test on the grounds of a lack of understanding. His Chinese Room thought experiment highlighted how behaviour may suggest understanding, but we cannot know. In this thought experiment Searle imagines he is *"locked in a room and given a large batch of Chinese writing"* (Searle, 1980) as well as possessing absolutely no knowledge of Chinese, he is, however, given a rule book that allows him to *"correlate one set of formal symbols with another set of formal symbols"* (Searle, 1980) and create responses to questions being posed by native Chinese speakers. Searle argues that the native speakers inputting questions, when they read the responses, would infer an understanding on Searle-in-the-room's part. But this is obviously wrong. In the same way, with a computer or AI we can never expect an understanding as we have of the material we give them. As such, even if a system passes the Turing Test, we cannot attribute understanding or comprehension. So to argue that an AI is thinking rationally and autonomously is false. This is reminiscent of the case of Koko the Gorilla who "mastered" sign language and could communicate with over 1000 signs. Owing to the fact that Koko's interpreter refused to allow scientific studying and was the sole communicator with Koko, there is some ambiguity about whether the Gorilla could be said to actually speak sign language. As an Economist article written shortly after her death wrote "*the fact that Koko could communicate should not mislead observers into thinking she possessed language"* (Johnson, 2018). In the same way that Koko could respond to stimuli with appropriate outputs, such as labelling a cat with the appropriate sign for "cat", an AI may be able to express even greater complexities, but an understanding of the output is fundamental. It is this understanding that seems to draw the line between an actual rational being, and a very good imitation.

Some would argue that AI will inevitably be able to possess the understanding of humanity if we are able to perfectly mimic the structure of the human brain, that is map all of the neurophysiology of the human brain onto a circuit level. This is a physicalist view of the mind, that mental states are identical to physical states. This has been addressed by many philosophers of mind and language over the last 100 years, but I shall put forward the arguments of John Searle and Jerry Fodor to suggest why this is incorrect.

Later in his paper on the Chinese Room, Searle used an adaptation of his previous thought experiment to refute this view of the mental being reduced purely to the physical in response to a critique. To parallel the firing of neurons, he argued that if you do away with the man in the room reading a rule book and writing symbols, and replace him with a man *"operat*[ing] *an elaborate set of water pipes with valves connecting them"* (Searle, 1980) of such complexity that it perfectly maps the firing of neurons in the brain, there is still no understanding in the system. It is an elaborate factory churning out answers. It is no more of a mind with understanding than your calculator. To Searle the crux of the issue is that *"it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states"* (Searle, 1980). Thus Physicalism, and any system whereby the brain is perfectly mimicked, fails. There must be something more that we possess as humans.

Jerry Fodor, taking a different stance, focused on the reductive nature of physicalism, and saw this as ludicrous. He employed a *reductio ad absurdum* argument to show that to reduce psychology to biology is implausible. Fodor argues that implicit in physicalism is a belief that "*special sciences must reduce to physical theories"* (Fodor, 1974), where special sciences are those other than fundamental physics, but being made up of physical particles, they are reducible to their simpler constituents. So, in the case of the mind, psychology is reduced to neurobiology, as anyone programming AI would have us believe. This produces an issue for Fodor when we take something like Gresham's Law in economic theory that "*says something about what will happen to monetary exchanges under certain situations"* (Fodor, 1974) specifically how the value of a currency relates to it being removed from circulation by a less valuable currency. He argues that *"any event which consists of a monetary exchange (hence any event which falls under Gresham's law) has a true description in the vocabulary of physics and in virtue of which it falls under the laws of physics"* (Fodor, 1974), hence we should be able to explain the value of currency and the shifts in trends if we have all of the physical data. This, Fodor argues, and many would agree, is absurd. Even if everything afflicted by Gresham's law (the currency, the banks, the bank tellers, the citizens) are made up of physical stuff, we have absolutely no reason to believe that a complete understanding of particle physics would transpose onto economics in any way. This view is wrong because of the "*wildly disjunctive"* (Fodor, 1974) nature of the special sciences – there are multiple ways of realising currency for instance, or in the case of brains, we have token physicalism (Fodor, 1974) the view that any mental state can be realised and is realised by an array of different physical states. As such to try and impose physics on any special science is spurious. In this sense special sciences are a unitary affair that, when you look deeper at their parts, become unfounded. That is, you can understand the psyche with psychology, and economical transactions with economics but to assume a hierarchy is to be mistaken. Hence, through Fodor's logic to assume the nature of the mind can be ascertained wholly physically is wrong.

Where this tangent leaves us is this – the nature of the programming of AI and their construction means that, while they may be able to perfectly mimic human activity, we cannot ascribe to them rationality and self-consciousness. Nor can we expect that by perfectly copying a human brain onto a computational system we will create a rational and self-conscious being. Hence the criteria we have given for moral status, the autonomy, reason and self-consciousness that we possess, is inapplicable to AI. Therefore, they are not beings entitled to moral status. But I wish to put forward that whether you agree with this or not is irrelevant – the entitlement to being treated moral and factoring into any moral calculation is owed to AI no matter what – moral status or not.

My next step in the argument is to try and tie these two sections together: why AI, given their lack of moral status, have moral worth regardless of the circumstance. Kant has two reasons for suggesting that animals be treated in a considerate way: 1) their similarity to humans in function and appearance warrants their good treatment, and following from this, 2) our treatment of animals is indicative of how we treat other humans. I would like to suggest that both of these arguments are also applicable to AI. First, based on their similarity to humans in appearance and function. This almost goes without saying, AI have, for the most part, been created to fill the roles of humans and more often than not imitate humans in their actions and features. Take, for instance, *Sophia* created by *Hanson Robotics* who was designed to look like the actress Audrey Hepburn - here we have an example of an AI modelled exactly off of a human, hence from Kant's logic based on their shared function and appearance, we ought to grant moral duties with respect to them.

Secondly, regarding what the treatment of such beings represents. If we take a look at Asimov's Bicentennial Man again, there is a passage in which this is cleverly portrayed. In this passage, the main character, the self-conscious AI Andrew, is being attacked by bullies who mercilessly attempt to make him take himself apart. When a friend of Andrew’s comes and attempts to stop their aggressions, Asimov writes the following “*They were smiling. The tall one said lightly, ‘what are you going to do, pudgy? Attack us?’”* (Asimov, 1976). Implicit in this passage we can see what Kant was eluding to. In the treatment of beings with which we can empathise, such as humanoid AI, or that we know feel pain, animals, there is a bleeding into the way we treat other humans. This is why we are so appalled by acts of animal abuse; normal humans empathise with their mistreatment and feel pity or are angered at such a prospect. In creating AI, be they self-conscious or not, it is within our human nature to wish for them to be treated well – just as is the case with animals. We know that a kitten cannot solve linear algebra, nor can it question its own existence, but we still demand it be treated well, and in doing so show our compassion as humans towards each other. As such it is only right that AI be treated in a moral fashion, and with their interests taken into consideration not out of duty to them or the categorical imperative but with respect to what they represent. This is a similar conclusion to the one reached by Anderson who argues “*we can require them to do things to serve our ends, but we should not mistreat them”* (Anderson, 2011).

This all suggests that philosophers have an incorrect view of moral rights in the way it was illustrated at the start. To ground moral treatment in some fictitious structure of moral status is wrong. There are so many different views one can take, it ends up becoming a straw man argument, no single view shall ever satisfy everything we feel is owed moral rights. As such I suggest that the best option would be to do away with the concept of moral status completely and focus on what it is that drives us to act compassionately towards others be they human or otherwise. It could be reason, but never does one say “I’m glad those firemen saved that infant from the blaze, seeing as he has the future capacity for intellect”, nor could it be sentience, as this leaves a hole in the reasoning when we think of something like the death of the Dog in the movie I am Legend, a scene that for many is immensely emotional. This character is fictional, and never existed as we perceived it other than as pixels on a screen. There was never any genuine harm, and yet, it is emotional enough to bring people to tears. This brings the idea of sentience into question: it is not that there was any actual pain inflicted that makes us upset, but the concept of suffering that leads us to empathise. I believe that the question of moral status is irrelevant. If the entity shares some trait or another with humans then we will undoubtedly empathise with it, such is our nature. So, if we encounter a being that shares sufficient attributes with humanity then it ought to have rights with respect to what it represents: an existing being. It is in features that entities share with us that we can fully understand them as existing as we do in some way. So, we can conclude that such an entity is owed rights, even if not for the same reason as we would give to other humans.

Thus, we see that the framework indicated at the beginning, of which Tibor Machan is one of the many proponents, is a false depiction. To base the applicability of moral rights on the crude idea of moral status is weak, owing to the various stances’ flaws. The theory is too full of holes to stand any ground, and so the best solution, as I see it, is to take Kantian view with regard to humanity and give moral rights to other beings with respect to their existence and similarities to us. In such a way AI, while lacking moral status on the grounds I have accepted, as well as many other definitions, are still owed moral rights, a significant fact to take into account given their rise in deployment today.

Bibliography

Alexander, L., & Moore, M. (2016, October 20). *Deontological Ethics*. Retrieved from Stanford Encyclopaedia of Philosophy: <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/>

Anderson, S. L. (2011). *The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics.* Cambridge: Cambridge University Press.

Asimov, I. (1976). *The Bicentennial Man and Other Stories.* Garden city, NY: Doubleday.

Bentham, J. (1823). *An Introduction to the Principles of Moral and Legislations.* London: Pickering.

Dabby, R. (2012). Pain disorders and erythromelalgia caused by voltage-gated sodium channel mutations. *Curr Neurol Neurosci Rep.*, 76-83. Retrieved from National Institute of Health : https://ghr.nlm.nih.gov/condition/congenital-insensitivity-to-pain#statistics

Fodor, J. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese October issue*, 97-115.

Johnson. (2018, July 5). *What Koko the gorilla could and couldn't do*. Retrieved from The Economist: https://www.economist.com/books-and-arts/2018/07/05/what-koko-the-gorilla-could-and-couldnt-do

Kant, I. (1780). *"Our Duties to Animals" in Lectures on Ethics (Infield, L. trans. 1963).* New York NY: Harper & Row.

Kant, I. (1785). *The Groundwork for the Metaphysics of Morals (Paton, H. J. trans. 1963).* New York NY: Barnes & Noble.

Kant, I. (1998). *Critique of Pure Reason (Guyer, P. & Wood, A. trans.).* Oxford: Oxford University Press.

Machan, T. R. (1991, Apr). Do Animals have Rights? *Public Affairs Quarterly Vol. 5 No. 2*, 163-173.

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosopchical Review Vol. 83 No. 4*, 435-450.

Reath, A. (2006, May). *Oxford Scholarship Online*. Retrieved from Agency and Autonomy in Kant's Moral Theory: Selected Essays: http://www.oxfordscholarship.com/view/10.1093/0199288836.001.0001/acprof-9780199288830-chapter-6

Ryder, R. (2010). Speciesism Again: the original leaflet. *Critical Society Issue 2 Spring*.

Searle, J. (1980). Minds, brains, and programs. *Behavioural and Brain Sciences*, 417-457.

Singer, P. (1995). *The Oxford Companion to Philosophy, edited by Ted Honderich.* Oxford: Oxford University Press.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind Issue 49*, 433-460.

Warren, M. A. (1997). *Moral Status - Obligations to Persons and Other Living Things.* Oxford: Clarendon Press.