18/09/2018

Contemporary Political Philosophy

2500 words

# The Talos Principle:
# A Defence of the Rights of Artificial Intelligence


## Abstract

This essay will discuss the point at which artificial intelligence will have rights that need to be recognised. The focus will be on Bentham, Gewirth, Locke, and in particular, Kant's socio-psychological properties, while also arguing that focusing on the biological component is a logically incoherent approach. It will then address three objections: the Hobbesian objection in 'otherness', Desjardins' non-identity argument, and the objection of eternal debt.


## Essay

The greek myth of Talos tells the story of a man of brass, created by Hephaestus to protect the island of Crete. He had "one blood vessel that extended from his neck to his ankle… [and when opened the blood] all flowed out and he expired."[1] He had wishes, beliefs, he spoke, and he moved of his own volition. It was Medea, a witch, who connived to drive him from sanity to madness.[2] With all this, did he not have the essential properties of a human? The obvious refutation to this is he was governed by the laws laid down in his design by Hephaestus; that Talos bears artificial limbs and an artificial mind. But then, if Man is governed by laws of nature, Man could be seen as a machine and a machine Man. If this is so, then does Talos not deserve to be recognised with all the rights of Man?

If two different beings are given a different moral status in society, then there must be some *conceptual distinction* between the two that permits that difference in moral status.

---

[1] Apollodorus *Bibliotheca* (circa 200AD) translated by Aldrich.
[2] Above, n1.

What is the 'conceptual distinction'?

In discussing the rights of artificial intelligence, we will assume a number of properties to them. The notion of 'human rights' denotes the property of being human, but Harris writes that this property is not conventionally a biological one.[3] Schwitzgebel and Garza hold that, "it shouldn't matter to one's moral status what kind of body one has".[4] While rights are often divided along the lines of species, this is not the only criteria. The liberal democracy regards race, sex, or sexual orientation as irrelevant when constructing a 'conceptual distinction' with rights.[5] Likewise, a human with an artificial leg or arm is not regarded as less-human by the law. Artificial hearts and manufactured blood even keep many alive. To expand on the thesis-metaphor, if this is so, then the brass arms, legs, and chest of Talos do not make him less human. While some critics point to the normative value of DNA, genome editing does not make one less human. On these collective grounds, it would be difficult to argue that 'artificially' created DNA was any less 'human'. Hephaestus' programming of Talos may be 'artificial', but it is no less human in its design.

Where the 'kind of body' does not matter, the conceptual distinction between humans and animals may offer some assistance. Schwitzgebel and Garza point to what they call the "social and psychological properties"[6] of a human, but give no definitive definition as to what these are, except that humans have them while animals do not. Mechanical differences matter only insofar as they manifest in different social and psychological properties. At the point artificial intelligence attains these abilities, there will be conceptual distinction that disentitles them from human rights.

---

[3] Harris, J *Enhancing Evolution: The Ethical Case for Making Better People* (Princeton University Press, Princeton, 2007).
[4] Eric Schwitzgebel and Mara Garza "A Defence of the Rights of Artificial Intelligences" (2015) 39(1) PSF at 104.
[5] Above, n4 at 104.
[6] Above, n4 at 103.

What are these 'psychological and social properties' that create the conceptual distinction?

**A Bentham Response**

Bentham's utilitarian maxim, underpinned by the "two sovereign masters, pain and pleasure"[7], emphasises one psychological property in the extension of rights: "...the question is not, Can they reason? nor, Can they talk? but, Can they suffer?"[8] However, this is a difficult distinction to draw, as animals feel pain, and yet we draw a conceptual distinction between them and humans. This alone could not elevate artificial intelligence to 'human-grade' properties.

There is also the very real possibility that humans would design artificial intelligence to *not* feel pain in the narrowest sense, such as the pain one feels from being stabbed. However, a broader understanding of pain, which we will call 'secondary pain', such as the emotions one feels when they suffer loss, betrayal, or unkindness would be an intrinsic part of the 'psychological and social properties' we will discuss forthwith.

**A Kantian Response**

Kant provides a more robust understanding of these properties with rational capacities. Rachels in *The Elements of Moral Philosophy* notes that rational beings are "capable of reasoning about [their] conduct and... freely decid[ing] what [they] will do."[9] This rational psychological property allows a being to consider the Categorical Imperative and *choose* to abide by it: that their maxim should apply universally. This is the ability to recognise what Kant calls the 'righteous law', which animals cannot do, and "because he has these capacities, a rational being is responsible for his actions."[10] Therefore, where a being is equally as free as a human,

---

[7] Jeremy Bentham *The Collected Works of Jeremy Bentham*, general editors: J. H. Burns, J. R. Dinwiddy, F. Rosen, T. P. Schofield (Athlone Press, Oxford, 1988) at 8.
[8] Jeremy Bentham *Introduction to the Principles of Morals and Legislation* (reprint: Athlone Press, 1970) at 144.
[9] James Rachels *The Elements of Moral Philosophy* (Random House, 1986) at 122.
[10] Above, n9 at 123.

and equally responsible, they must be equally entitled to the same rights that come with those responsibilities.

This is complemented by Gewirth's argument that 'human' rights should be based on human agency. The conceptual distinction is drawn where a being can make a 'prudential rights claim' for themselves against others.[11] This is the psychological property: They have the ability to recognise the "indispensable conditions of agency and action as necessary goods".[12] It also requires a social property: They have the ability to relate to others and empathise with them in such a way they can recognise this. If capable of making this rights claim, reciprocity would require it returned them.[13] Within this conceptual distinction is Bentham's secondary pain, where the artificial intelligence could understand the unfairness and feeling of betrayal that would come with not being equally respected, despite being equally rational.

**A Lockean Response**

Kant's rational psychological capacities can be paired with Locke's consent theory, which places more emphasis on the social properties of a being when it comes to the recognition of rights. While Locke draws his natural right that "every Man has a Property in his own Person" [14] largely from divine tradition - a tradition absent in the manufactured creation of artificial intelligence - he also places emphasis on a person being one that can use "reason and reflection, and can consider itself as a self."[15] This alludes to a conceptual distinction of consciousness, which an artificial intelligence would have.

With these properties, a being can attain certain rights *against* others by social contract with them. Should an artificial intelligence enter into social contract with humanity, it would be entitled to same the same human rights, This requires the social properties to engage with others

---

[11] Alan Gewirth *Human Rights: Essays on Justification and Applications* (Chicago University Press, Chicago, 1982) at 68.
[12] James Nickel "Human Rights" (2014) Stanford Encyclopedia of Philosophy <https://plato.stanford.edu/entries/rights-human/>.
[13] Above, n11 at 73.
[14] John Locke *Two Treatises of Government* (Awnsham Churchill, London, 1689) at 194.
[15] John Locke *An Essay Concerning Human Understanding* (2nd ed, Clarendon Press, Oxford, 1975) at 335.

and the ability to do so with their own interests in mind. These would be the protection of their life, liberty, and property that Locke recognises from being able to "consider itself as a self." The focus here is on their ability to integrate into a new society. Dennett's work elaborates on the mechanical nature of consciousness, that we are our minds and our minds are "robots made of robots made of robots."[16] A parallel can obviously be drawn from this in that we, as people, may all be seen as 'artificial' in this light, if the only 'real' thing about consciousness is the subjective element.

Talos suffered in dying when "an arrow from Poeas in the ankle finished him",[17] but this psychological property of Bentham's is insufficient. Instead, it is his capacity to wish, to speak, and to hold his own beliefs that brings him up to our Kantian conceptual distinction. For if he could be driven mad by Medea, then he must have once been rational. He considered himself as a self, and it is in this vein he could have made a prudential rights claims. No conceptual distinction could be drawn that would demand a different moral status. What could he be, then, if not a Man?

The first objection: Hobbes

Hobbes's knockdown argument is that no rights can exist in the State of Nature which is "a man against every man."[18] The only 'true libert[y]' is that to defend oneself to whatever ends needed, and all other rights are subservient to that. Within this, humanity would exist in a state of war with artificial intelligences, who are not party to any social contract, and we would thus be justified in oppressing them in any way whatsoever. In the same fashion, we would owe no rights to an alien who comes to Earth.[19]

The answer to this is in Hobbes' theory of concern itself: Across his work, Hobbes highlights the social properties of obligations and the 'true liberties of subjects'. These include

---

[16] Daniel Dennett *Consciousness Explained* (United States, Little Brown & Co, 1991) at 197.
[17] Above, n1.
[18] Thomas Hobbes *Leviathan or The Matter, Forme and Power of a Common-Wealth Ecclesiasticall and Civil* (London, 1651) at 77.
[19] Above, n4 at 16.

the right to refuse to incriminate themselves or loved ones, and the right to defend not just themselves, but their honour and those they love.[20] Under this model, the obligations of the individual are to family first, community second, and others third. The misconception is that artificial intelligence would fit into 'others' where Hobbes took a realist approach to international relations. They would, in fact, be more akin to family. As established, the biological nature of humanity, here being that of a natural birth, is largely irrelevant. Talos, being brought into existence through the design and will of Hephaestus, is more akin to his son and creation than an 'other'. In this vein, artificial intelligence would be a child of humanity, born beneath the thumb of the Leviathan. They are not outside it, but under its care.

To argue otherwise would suggest emphasis should be placed on either (a) the way in which an individual is brought into the world or (b) how humans feel about artificial intelligence personally. The former would bring into question the familial relationship of those born via IVF treatment or fetuses undergoing genome editing, while the latter is a subjective test that has been used to alienate other races and sexes from rights before.

The second objection: Desjardins' 'non-identity' problem

Hephaestus brought Talos into existence with the intention of him defending the island of Crete. Without this purpose in mind for his creation, Talos would not exist. Despite Patrick Henry's famous platitude, "Give me liberty, or give me death!",[21] if we are to assume Gewirth's "agency and action [are] necessary goods",[22] it logically follows existence itself must be the most fundamental necessary good. Talos could not complain where he enslaved to defend Crete against his will if the alternative is non-existence. This is also a partly Hobbesian argument, where the it is better to live an oppressed life in the Civil State than exist in the State of Nature, where one's life is forfeit.

[20] Above, n18.
[21] Patrick Henry (St. John's Church Richmond, Second Virginia Convention, March 23 1775).
[22] Above, n11.

Very rarely do humans create without intent for application. Artificial intelligences are not usually seen as 'ends in themselves',[23] as Kant would have it, but as a means to more efficient regulation of shipping lines, companionship for the elderly, or replication of a genuine relationship of love with a female for a sexually deprived male. Without such a purpose, the artificial intelligence would not exist. Desjardins provides the logical conclusion to this: "... it makes little sense to say that they would be 'better off' if we had made the other choice. Because different . . . decisions result in different [beings]."[24] Because of this, living with one's existence being conditional on diminished rights must be better than non-existence.

Schwitzgebel and Garza raise this argument in respect of animals whose "death for meat is conditional of its existence",[25] but this argument fails when it comes to people with the social properties we have discussed prior. As established, the relationship between artificial intelligence and human creator is more akin to parent-child than anything else. Schwitzgebel and Garza argue that, "although the child in some sense "owes" its existence to [their parents], that is not a callable debt."[26] Beyond this, a society that gives a different moral status to two different beings without that conceptual distinction would make the distribution of rights arbitrary.

The third objection: eternal debt

Would Hephaestus have created Talos if he knew he would have to work an hour every day to create the ichor that ran in his veins as blood? Arguably not. It will be a considerable financial investment to develop artificial intelligence. Once created, there will be expenses to sustain them - if only in the cost of power, renting a premises, or repairing their android body, should they have one. It is possible the creators would not have brought them into existence if they knew they would be obligated to pay these expenses in perpetuity. This argument follows

[23] Immanuel Kant *Groundwork of the Metaphysics of Morals* Translated by Mary J Gregor (Cambridge University Press, Cambridge, 1991).
[24] Joseph R Desjardins *Environmental Ethics,* (Wadsworth Publishing Company, Belmont, CA, 1993) at 78.
[25] Above, n4 at 18.
[26] Above, n4 at 19.

similar to Desjardins' argument, but inverts it in that the concern is that the condition applies to humanity. Underpinning this is also the notion that being property diminishes one's rights.

The artificial intelligence shares a parental relationship with its creators, being either born of design or accident through no choice of its own, inheriting whatever its parts may be from its creator, but crucially - with human-grade capacities. As discussed, the biological component to their birth is not required for this relationship. Because of this, we can use Archard's commonly accepted argument of special obligations for parents: "[T]hose who cause children to exist thereby incur an obligation that they are adequately cared for".[27]

While some would argue for explicit voluntarism, it would be inconsistent for a society to hold to what Dworkin called "associative obligations… attache[d] to membership in some biological or social group"[28] with regards to human parental relationships, but not this relationship. Human rights are not just negative, but also comprise of natural duties we owe to others - such as to children. The extent of this is unclear, as these special obligations decrease as children become sufficient. Yet, an artificial intelligence may be created fully sufficient. However, the costs of merely being sustained must come within this.

This is also complicated by one other situation: Dworkin's "associative obligations" do not extend to the same extent to unborn children who have the potential for the psychological and social properties we conventionally use as the threshold. They have no right to life. It is unclear if a program that could, either definitely or only probabilistically, develop the social-psychological 'human-grade' properties would be owed the same obligations. As Hephaestus should care for Talos, we are bound to our technological children.

**Conclusion**

No conceptual distinction that would allow for a different moral status could be drawn between an artificial intelligence and a human if they attain the same socio-psychological properties. Rights arise here not from the capacity for pain and pleasure, but Kantian rationalism

---

[27] David Archard *Procreation and Parenthood: The Ethics of Bearing and Rearing Children* (Oxford University Press, Oxford, 2010) at 127.
[28] Ronald Dworkin *Law's Empire* (Harvard University Press, Harvard, 1986) at 196.

- the ability to consider the Categorical Imperative, the "righteous law", and abide by it, understanding all the facets this involves. It is then the necessary good of agency that gives rise to a prudential rights claim. Being equally free and equally subject demands equal rights. The Hobbesian objection that excludes rights based on being in a state of war with humanity wrongfully places emphasis on the biological element of artificial intelligence, when in reality, the relationship is more akin to a parent-child. As persuasive as Desjardins' non-identity argument may be, it would be logically inconsistent to apply this conditional-existence moral status to artificial intelligence but not to children where no conceptual distinction can be drawn. Finally, the responsibility for the 'eternal debt' cannot be imputed to a being who had no choice in their existence. Under any model of special obligations, it would fall to the 'parent', who either tacitly, voluntarily, or through Dworkin's "associative obligations" took up this obligation. If we are Hephaestus, and if Talos is a Man, and Man has rights, then it is our duty to recognise them, for he is no more artificial than you or I.

**Bibliography**


Alan Gewirth *Human Rights: Essays on Justification and Applications* (University of Chicago Press, Chicago, 1982).

Apollodorus *Bibliotheca* (circa 200AD) translated by Aldrich.

Daniel Dennett *Consciousness Explained* (United States, Little Brown & Co, 1991) at 197.

David Archard *Procreation and Parenthood: The Ethics of Bearing and Rearing Children* (Ocford University Press, Oxford, 2010) at 127.

Eric Schwitzgebel and Mara Garza "A Defence of the Rights of Artificial Intelligences" (2015) 39(1) PSF 98-115.

Harris, J *Enhancing Evolution: The Ethical Case for Making Better People* (Princeton University Press, Princeton, 2007).

Immanuel Kant *Groundwork of the Metaphysics of Morals* Translated by Mary J Gregor (Cambridge University Press, Cambridge, 1991).

James Rachels *The Elements of Moral Philosophy* (Random House, 1986) at 122.

Jeremy Bentham *Introduction to the Principles of Morals and Legislation* (reprint: Athlone Press, 1970) at 144.

Jeremy Bentham *The Collected Works of Jeremy Bentham*, general editors: J. H. Burns, J. R. Dinwiddy, F. Rosen, T. P. Schofield (Athlone Press, Clarendon Press, Oxford, 1988).

James Nickel "Human Rights" (2014) Stanford Encyclopedia of Philosophy <https://plato.stanford.edu/entries/rights-human/>.

John Locke *An Essay Concerning Human Understanding* (2nd ed, Clarendon Press, Oxford, 1975) at 335.

John Locke *Two Treatises of Government* (Awnsham Churchill, London, 1689).

Joseph R Desjardins *Environmental Ethics,* (Wadsworth Publishing Company, Belmont, CA, 1993) at 78.

Mark Coeckelbergh "Robot rights? Towards a social-relational justification of moral consideration" (2010) 12(3) EIT 209-221.

Mathias Risse “Human Rights and Artificial Intelligence: An Urgent Agenda” (2018) 4(1) RP 1-16.

Patrick Henry (St. John’s Church Richmond, Second Virginia Convention, March 23 1775).

Ronald Dworkin *Law’s Empire* (Harvard University Press, Harvard, 1986) at 196.

Thomas Hobbes *Leviathan or The Matter, Forme and Power of a Common-Wealth Ecclesiasticall and Civil* (London, 1651).