The hard limit on human nonanthropocentrism

Michael R. Scheessele

Indiana University South Bend

Department of Computer and Information Sciences

South Bend, IN USA

Email: mscheess@iusb.edu

ORCID: 0000-0002-0014-9353

Acknowledgements

I wish to thank Indiana University South Bend for the Faculty Research Grant that supported this research. I especially want to thank Mahesh Ananth for many lively discussions and valuable suggestions on several drafts of this paper.

Declarations

Funding

This research was funded by an Indiana University South Bend Faculty Research Grant.

Conflicts of interest/Competing interests

The author declares that he has no conflict of interest.

Availability of data and material

Not applicable.

Code availability:

Not applicable.

Authors’ contributions

Sole-authored paper.

The hard limit on human nonanthropocentrism

Abstract

There may be a limit on our capacity to suppress anthropocentric tendencies toward non-human others. Normally, we do not reach this limit in our dealings with animals, the environment, etc. Thus, continued striving to overcome anthropocentrism when confronted with these non-human others may be justified. Anticipation of super artificial intelligence may force us to face this limit, denying us the ability to free ourselves completely of anthropocentrism. This could be for our own good.

Keywords anthropocentrism moral status intelligent machines super artificial intelligence

1 Introduction

To many philosophers, anthropocentrism is a polysyllabic dirty word. For example, Gunkel (2012) and Rae (2016) describe critiques implicating the role of anthropocentrism in sexism and racism. Others challenge anthropocentrism by their efforts to extend the scope of moral concern beyond humanity. Accordingly, animal (Singer 1974; 1985; Regan 1985), environmental (Taylor 1981/2010; Leopold 1949/1977/2010), and information ethicists (Floridi 2008; Floridi and Sanders 2004) argue for moral consideration of non-human entities. Given continuing advances in artificial intelligence, some even insist that we could have moral obligations to intelligent machines (Gunkel 2012; Scheessele 2018; Tavani 2018; Gerdes 2015; Coeckelbergh 2010).[1] Even so, some worry about a continuing destructive impact of anthropocentrism (Gunkel 2012). An underlying concern is that anthropocentrism devalues a range of non-human entities.

There may be a limit on our capacity to suppress anthropocentric tendencies toward non-human others, though. Normally, we do not reach this limit in our dealings with animals, the environment, etc. Thus, continued striving to overcome anthropocentrism when confronted with these non-human others may be justified. Achievement of super artificial intelligence (super AI), however, may force us to face this limit, denying us the ability to free ourselves completely of anthropocentrism. Given concerns that super AI could pose an existential threat to humanity (Bostrom 2014/2016), this might be for our own good.

In recognition of ongoing advances in artificial intelligence, *machine ethics* (Anderson and Anderson 2007; Wallach and Allen 2009) has focused on how intelligent machines *should* act toward humans in a variety of morally charged contexts. This approach is agent-based and anthropocentric. Gunkel (2012) observes that, in Western philosophy, a moral agent not only has obligations to others, but also others have obligations to the moral agent. With respect to intelligent machines, Gunkel illuminates the disparity between the focus on obligations of an intelligent machine toward us (i.e., a moral agent-based perspective) and on our potential obligations toward an intelligent machine (i.e., a moral patient-based perspective). Since then, attention to the question of moral status for intelligent machines has increased (Gunkel 2014; Gerdes 2015; Scheessele 2018; Tavani 2018), coupled with ever-louder calls for moral and legal robot rights (Prodhan 2016; Gunkel 2014; 2018a; 2018b). Anthropocentrism,

[1] Further, debate over legal personhood for robots has crossed over into the mainstream (Prodhan 2016; Floridi and Taddeo 2018).

according to thinkers such as Gunkel (2012), is a chief obstacle to the determination of moral status of, and extension of rights to, intelligent machines.

It may be defensible to decry anthropocentrism when considering non-machine others. Even intelligent machines[2] could have moral status. Still, the case of at least certain intelligent machines may be different from the previous cases of non-humans, such as animals and the environment. In the case of super AI, for example, we may confront a limit beyond which anthropocentrism is unavoidable. If correct, we ought to consider this now, prior to the advent of super AI, in discussions of the moral status, as well as moral and legal rights, of intelligent machines.

The organization of this essay is as follows. In the next section, I briefly review the concept of anthropocentrism and nonanthropocentric challenges to anthropocentrism. Following that, I suggest a hard limit on human nonanthropocentrism when faced with the prospect of super AI. After that, I consider some possible objections to my conjecture of a hard limit. In the final section, I draw some general conclusions related to my defense of a circumscribed anthropocentrism.

2 Anthropocentrism and its Nonanthropocentric Challenges

Thompson (2017) observes that "anthropocentrism is commonly understood as a theory of value which maintains that only human beings[3] or their experiential states have intrinsic moral value." Thompson notes that "there are at least three forms of anthropocentrism discussed in the literature": ontological anthropocentrism, ethical anthropocentrism, and conceptual anthropocentrism. *Ontological anthropocentrism* "locates humans at the center of creation, as the end or reason for which everything else in the material world exists." *Ethical anthropocentrism* is either "strong" or "weak." Per Thompson, "strong ethical anthropocentrism" maintains that nothing other than humans have intrinsic moral value, whereas "weak ethical anthropocentrism" allows for entities other than humans to have intrinsic moral value, but less than that of humans. In either the strong or weak version, human intrinsic moral value "trumps" other entities, in terms of their value. According to Thompson, ontological anthropocentrism can stand in support of ethical anthropocentrism. However, it does not entail ethical anthropocentrism. As

[2] "Intelligent machine," as used here, refers to a reasonably sophisticated product of artificial intelligence (AI) or related disciplines. An intelligent machine may be prominently hardware, as with a robot or digital computer, or primarily software, as with a virtual agent or software-based system. An intelligent machine may stand alone or be embedded in another artifact. It may be silicon-based or not—as with products from the field of synthetic biology. Further, an intelligent machine, for purposes of this essay, could be a hybrid of two or more materials—a silicon-based digital computer interfaced with a neural circuit made from biological material, perhaps.

[3] Although some may wish to stretch the meaning of "human being," here I use "human" and "human being" to refer to a member of *Homo sapiens*.

Thompson points out, ethical anthropocentrism has been justified separately on the basis of qualities such as consciousness, rationality, moral agency, etc. Thus, modern "scientific worldviews," which tell against ontological anthropocentrism, do not necessarily threaten ethical anthropocentrism.

Conceptual anthropocentrism, according to Thompson (2017), "is the idea that human beings can only comprehend the world from a characteristically human perspective—from within a human conceptual framework." Thompson elaborates:

> For environmental ethics, the most significant form of conceptual anthropocentrism is not theoretical but practical: limits imposed by the structure of human normative and axiological capacities. However we appreciate value or whatever we value, our valuing must always done [sic] from a human perspective. "All human values are human values," Minteer explains, "including the intrinsic value that ethical nonanthropocentrists ascribe to nature" (Minteer, 2009). The bearing this has for environmental ethics is fundamental and inescapable but often overlooked. (p. 79)

My conjecture of a "hard limit" on human nonanthropocentrism is squarely a form of conceptual anthropocentrism. Echoing Thompson, I maintain that there are "[practical] limits imposed by the structure of human normative and axiological capacities." My conjecture is that limits on *nonanthropocentrism* entailed by *conceptual anthropocentrism* will become obvious in the face of super artificial intelligence (in a way they have not yet become obvious), should super AI occur (or seem to become more likely to occur).

By "human[4] nonanthropocentrism," I mean the "ethical nonanthropocentrism" characterized by Thompson (2017) as follows:

> If strong ethical anthropocentrism is the view that *only* human beings bear an intrinsic moral value or are morally considerable, then ethical nonanthropocentrism implies that some part or parts of the nonhuman world are intrinsically valuable. If weak ethical anthropocentrism is the view that the intrinsic moral worth of human beings *always overrides* the moral significance of nonhumans, then ethical nonanthropocentrism implies that the moral value of human beings does not always trump all other values. (p. 87, endnote 6)

[4] The "human" qualifier is necessary if one imagines an intelligent machine designed not to regard humans as the only entities with intrinsic moral value.

In the remainder of this section, I provide several examples of nonanthropocentric challenges to anthropocentrism. I conclude with some critique of such challenges.

2.1 Challenges to Anthropocentrism

According to Thompson (2017), ethical nonanthropocentrism theories fall into one of three categories: sentiocentrism, biocentrism, and ecocentrism. To these, I add a fourth category, ontocentrism, which captures the Information Ethics (IE) of Floridi (Floridi 2008; Floridi and Sanders 2004). Examples of ethical nonanthropocentrism theories are provided next. The list of examples is not complete, and the brief description of each example is not intended to be comprehensive. Rather, the (hopefully representative) focus here is to provide a glimpse into how ethical nonanthropocentrism theories have extended moral consideration beyond human beings, thereby challenging anthropocentrism. Finally, I describe the “radical” nonanthropocentrism of Gunkel (2012) and Coeckelbergh (2012).

2.1.1 Sentiocentrism

In the 1960’s-1970’s, particularly in the U.S. and Great Britain, movements for increased women’s rights and civil rights fostered an environment more conducive to discussion and debate regarding moral consideration of previously excluded others (DeGrazia 2002). This gave rise to at least two prominent, but different, proposals for the moral consideration of animals.

In the first proposal, Singer (1974; 1985) marks sentience—the ability to feel, perceive, or experience suffering—as the threshold for moral consideration of an entity. If an entity is sentient, then its interests ought to be considered. Human and other vertebrate animals are sentient. So, too, are some invertebrate animals. Other invertebrates are not sentient (DeGrazia 2002). Therefore, we have no moral obligations to them, according to Singer (1974; 1985). Similarly, plants and trees, as well as machines and other artifacts, make no moral claims on us in Singer’s view.

Singer’s (1974; 1985) theory requires equal consideration of interests for all sentient animals, whether human or non-human, in utilitarian calculations of the greatest good for the greatest number. For Singer, equal consideration of interests does not mean that non-human animals are equal to humans in their capacities (e.g., intellect, moral capacity, etc.). It also does not mean that non-human animals should have the same rights that we do. As an example, he observes that a pig would have no use for the right to vote (Singer 1974). In a given situation, even the interests of a sentient, non-human animal may not be equal to those of a human. Singer (1985) uses lethal scientific experimentation in order to make his point. He asks the reader to assume that humans could be kidnapped from public parks for use in this sort of experimentation. He compares this to the case of non-human animals, plucked from their habitats for the same purpose. Assuming a human and non-human animal would suffer equally from the actual incarceration and experimentation, Singer argues how the human's suffering may be greater still, due to the dread a human would experience, while still free, from the prospect of being snatched for use in this type of experimentation. The presumption is that a non-human animal would not be aware that it could be taken suddenly for such a purpose. So, the non-human animal would not suffer dread prior to capture. In this case, Singer believes the non-human animal should be preferred to the human for such experimentation, because the human's suffering would be greater than the non-human animal's suffering, making interests of the human greater than those of the non-human animal. Singer finds this scenario acceptable because the choice of the non-human animal over the human would not be because the human is human. That is, the decision would not rest on speciesism, the preference for one species, in this case *Homo sapiens*, over another. Rather, the decision would be grounded in a utilitarian calculus with sentience as its foundation. Although speciesism is not equivalent to anthropocentrism (Faria and Paez 2014),[5] in this scenario the two would overlap, if the human were spared merely for being human.

The animal rights theory of Regan does not draw the line at sentience, but rather, at being "the experiencing subject of a life." This entails being "a conscious creature having an individual welfare" which matters to it, no matter the creature's value to others (Regan 1985). Being "the experiencing subject of a life" is the basis of an entity's "inherent value." While Singer's theory (1974; 1985) requires equal consideration of interests, Regan's theory, inspired by Kant, demands that "inherent value, then, belongs equally to those who are the experiencing subjects of a life"—no matter whether human or non-human animal (Regan 1985). In their calls for equality, whether in consideration of interests or in inherent value, between human and non-human animals, the theories of Singer and Regan, respectively, do not privilege humans over non-human animals. In this regard, both theories directly challenge ethical anthropocentrism, both strong and weak.

2.1.2 Biocentrism

Biocentrism proposes that life is the litmus test for moral consideration. Biocentrism was rigorously grounded philosophically in the late 20th century, largely due to the work of Paul Taylor (DesJardins 2015). The basis of

[5] For example, one could favor a particular non-human species over another species (human or non-human).

Taylor's (1981/2010) biocentrism[6] is an attitude of "respect for nature," where all living things have inherent worth. Taylor's theory also has a Kantian flavor. He notes the parallel of "respect for nature" to a Kantian attitude of "respect for persons." Taylor's concept of "inherent worth," while reminiscent of Regan's (1985) concept of "inherent value," differs in its foundation. While an animal's inherent value is grounded by being an "experiencing subject of a life," a natural entity's being alive and having "a good of its own" support its inherent worth. For an entity to have "a good of its own" means that "without reference to any *other* entity, it can be benefited or harmed" (Taylor 1981/2010). Taylor claims that a living entity can have a good of its own, even in the absence of sentience.

An attitude of respect for nature is universalizable and disinterested, according to Taylor (1981/2010). A moral agent with this attitude has dispositions such that the moral agent can be expected to adhere to duties, such as nonmaleficence and noninterference, and to develop character standards, such as fairness and benevolence, that facilitate promotion and protection of the good of living things. This leaves the question of "why *should* moral agents regard wild living things as possessing inherent worth?" The answer to this question, for Taylor, lies in the justification of the adoption of the attitude of respect for nature by all moral agents.

The basis of the attitude of respect for nature is a belief system which Taylor (1981/2010) calls "the biocentric outlook on nature." The biocentric outlook consists of four parts:

> (1) Humans are thought of as members of the Earth's community of life, holding that membership on the same terms as apply to all the nonhuman members.
>
> (2) The Earth's natural ecosystems as a totality are seen as a complex web of interconnected elements, with the sound biological functioning of each being dependent on the sound biological function of the others. …
>
> (3) Each individual organism is conceived of as a teleological center of life, pursuing its own good in its own way.
>
> (4) Whether we are concerned with standards of merit or with the concept of inherent worth, the claim that humans by their very nature are superior to other species is a groundless claim and, in the light of elements (1), (2), and (3) above, must be rejected as nothing more than an irrational bias in our own favor. (p. 518)

[6] There are other versions of biocentrism besides Taylor's version (Thompson, 2017, pp. 80-81). Thompson also points out that Taylor's biocentrism is *individualistic*, whereas *holistic* biocentrist theories are concerned with set(s) of living objects, such as species.

Taylor claims that this biocentric outlook can be used to justify the adoption of the attitude of respect for nature by all moral agents. It is beyond the scope of this essay to evaluate that claim; however, it is clear from the belief system above, which grounds the attitude of respect for nature, that Taylor's theory challenges anthropocentrism.

2.1.3 Ecocentrism

Ecocentrism is "the view that a collection of living and non-living things, which can be identified by compositional and functional characterizations and exhibit resilience as such, constitutes an ecosystem, and ecosystems are the loci of intrinsic value and hence moral significance" (Thompson 2017, p. 81). Ecocentrism is a holistic theory, with Leopold's (1949/1977/2010) "land ethic" as an early example, and with the longstanding theories of Callicott and Rolston continuing to serve as prominent versions of ecocentrism (Thompson 2017, p. 81).

Callicott (1984) admires Leopold's (1949/1977/2010) "land ethic" that "a thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise" (p. 504). As Callicott (1984 p. 303) observes, however, Leopold needs to defend this moral principle. Callicott's belief is that an ecocentric theory must be "at once humanistic, but not anthropocentric" (p. 304). Thompson (2017, p. 81) highlights the subjectivity of Callicott's theory by noting that "all attributions of intrinsic value are *anthropogenic*: originating in and dependent upon human acts of evaluation." Although this may not sound nonanthropocentric, Callicott (1984) reasons as follows. First, he adopts Hume's value theory, observing that, for Hume, "all behavior is motivated by passion, emotion, feeling, or sentiment" (p. 304). Callicott explains that Darwin took up Hume's position, "since emotion or passion is a more primitive and universal animal capacity than reason or any other supposed well-spring of moral behavior" (p. 304). Next, Leopold included Darwin's "origins of ethics" theory into his land ethic (p. 305). Despite the subjectivity of Hume's concept of value, Callicott concludes: "Therefore, the Darwin-Leopold environmental ethic, grounded in the axiology of Hume, is genuinely and straightforwardly non-anthropocentric, since it provides for the *intrinsic* value of non-human natural entities" (p. 305). He acknowledges, however, that "intrinsic value ultimately depends upon human valuers" (p. 305).

Rolston's ecocentric theory also specifies the intrinsic value of natural entities beyond humans. But this value is *found* in such entities, not merely ascribed to them by humans (Rolston 1975, p. 103). As Thompson (2017, p. 81) puts it, "value is discovered in, not conferred upon, nature and thus this is a theory of value that is objective and nonanthropogenic."

Hargrove (1992) claims that neither Callicott's nor Rolston's theories are free from anthropocentrism, however. Regarding the former theory, Hargrove believes that Callicott's "subjectivist position," particularly given his (Hargrove's) own belief that non-human creatures are also capable of valuing, prevents the theory from being nonanthropocentric. Hargrove likens Callicott's theory to "a slightly stronger (weak[7]) anthropocentrism than my own" (p. 196).

In regards to the theory of Rolston, Hargrove (1992) notes that "Rolston's nonanthropocentrism is also infected with anthropocentrism…, for he argues, against the biocentrism of Taylor, that humans are superior to the rest of nature and deserve special consideration, a *strong* anthropocentric claim" (p. 202). Beyond this surface observation, however, exists a deeper reason why Hargrove believes that Rolston's ecocentrism theory does not escape anthropocentrism. Hargrove posits that there is nonanthropocentric intrinsic value *and* anthropocentric intrinsic value.[8] Hargrove believes that anthropocentric intrinsic value rallies us with respect to the nonanthropocentric intrinsic value that we discover in nature:

> I want to argue here that anthropocentric intrinsic value judgments, rather than being in competition with nonanthropocentric intrinsic values, are absolutely essential if humans are to muster any environmental concern about nonhuman living centers of purpose (as well as many other natural entities) objectively existing out in the world. (p. 188)

The takeaway of this criticism is that Hargrove, unlike objectivist nonanthropocentrists (such as Rolston), advocates for more than one type of intrinsic value (pp. 187-188). For Hargrove, not only does anthropocentric intrinsic value complement nonanthropocentric intrinsic value by generating "moral behavior on behalf of the [living] creature" (p. 191), anthropocentric intrinsic value also allows for non-instrumental valuing of non-living entities (p. 192).[9] This

[7] The "weak anthropocentric intrinsic value" of Hargrove (1992) is not the same thing as the "weak [ethical] anthropocentrism" of Thompson (2017). Hargrove observed that "anthropocentric" was being used incorrectly as a synonym for "instrumental" (pp. 183-184) His use of "weak anthropocentric" implies anthropocentric value that need not be instrumental—value could be intrinsic instead. As with Thompson's use of "weak [ethical] anthropocentrism," Hargrove allows intrinsic value for some non-humans. With Hargrove, however, it is not clear (to me) that human intrinsic value always beats non-human moral value, as with Thompson's definition of "weak [ethical] anthropocentrism."

[8] Thompson (2017) clarifies Hargrove's nonanthropocentric intrinsic value by offering, as an example, the "good of its own" of a living organism. As an example of anthropocentric intrinsic value, Thompson observes how parents (non-instrumentally) value the life of a child for its own sake (pp.82-83).

[9] Hargrove's (1992, p. 192) interest is in protecting caves.

latter point has obvious relevance to the question of whether a non-living intelligent machine could have moral status grounded in intrinsic value.

2.1.4 Ontocentrism (e.g., Information Ethics)

A prominent version of information ethics (IE) is due to Floridi; a detailed description of his theory can be found in Floridi (2008) and Floridi and Sanders (2004). His theory is ontocentric, meaning that existence, or being, is of central importance in the theory. (By contrast, recall that life, or being alive, is the criterion in biocentrism.) Floridi's theory extends moral consideration to anything that exists, whether currently, in the past, or in the future. It also extends moral consideration to "ideal, intangible, or intellectual objects," rather than to merely physical objects (Floridi 2008). Further, it is a "patient-oriented" theory, like the various animal ethics, biocentric, and ecocentric theories mentioned above. This means that emphasis is placed on the recipient of moral action, rather than on the moral agent. As with most of the theories described previously, Floridi (2008) appears not to privilege humans over other entities, as evidenced by the "ontological equality principle":

> … IE holds that every entity, as an expression of *being*, has a dignity, constituted by its mode of existence and essence (the collection of all the elementary proprieties that constitute it for what it is), which deserve to be respected (at least in a minimal and overridable sense) and, hence, place moral claims on the interacting agent and ought to contribute to the constraint and guidance of his ethical decisions and behaviour. This ontological equality principle means that any form of reality (any instance of *information/being*), simply for the fact of *being* what it is, enjoys a minimal, initial, overridable, equal right to exist and develop in a way that is appropriate to its nature. (p. 48)

A non-human entity's "minimal, initial, overridable, equal right to exist and develop in a way that is appropriate to its nature" can be construed as a challenge to anthropocentrism.

2.1.5 Other-Oriented/Relational Challenge

The "other-oriented/relational challenge"[10] can be characterized by Gunkel's (2007) "thinking otherwise" and Coeckelbergh's (2012) "relational turn." The following brief descriptions of their work, as well as their

[10] Coeckelbergh and Gunkel (2016) have referred to this as "a relational and other-oriented concept" after the former's "relational turn" (Coeckelbergh 2012) and the latter's "thinking otherwise" (Gunkel 2007). I have just shortened their terminology to the *other-oriented/relational challenge*.

collaboration, are not intended to be comprehensive. Again, the (hopefully representative) focus here is to demonstrate how the other-oriented/relational challenge critically responds to anthropocentrism.

Gunkel (2012) welcomes the increasing inclusion of non-human others for moral consideration, as well as the shift from a moral agency to a moral patiency perspective entailed by the progression of ethical nonanthropocentrism theories. Yet, he believes the underlying approach of these theories is flawed. Gunkel (2014) comments on what he sees as the flaw in this "centrist approach" (e.g.,sentiocentrism, biocentrism, ontocentrism):

> All of these innovations, despite their differences in focus, employ a similar maneuver and logic. That is, they redefine the center of moral consideration in order to describe progressively larger circles that come to encompass a wider range of possible participants. Although there are and will continue to be considerable debates about what should define the center and who or what is or is not included, this debate is not the problem. The problem rests with the strategy itself. In taking a centrist approach, these different ethical theories (of which IE would presumably be the final and ultimate form) endeavor to identify what is essentially the same in a phenomenal diversity of different individuals. Consequently, they include others by effectively stripping away and reducing differences. This approach, although having the appearance of being increasingly more inclusive, effaces the unique alterity of others and turns them into more of the same. (p. 124)

Thus to Gunkel, the various ethical nonanthropocentrism theories use a strategy which is exclusive and which reduces the differences of a previously excluded other to sameness, when it finally includes that other. Gunkel crystallizes this point as follows:

> Exclusion is a problem because it calls attention to and fixates on what is different despite potential similarities. Inclusion is a problem, because it emphasizes similarities at the expense of respecting important differences. (Gunkel 2007, p. 174)

For Gunkel, exclusion and inclusion "are two sides of one coin."[11] He proposes a third alternative, "thinking otherwise," which is inspired by the continental philosopher Emmanuel Levinas.[12]

The work of Levinas, as with that of other poststructuralists, attempts to undermine and overthrow various binary oppositions (Gunkel 2018b). One such binary opposition is that of inclusion into or exclusion from the

[11] See Gunkel (2012) for a similar view on the moral agency/moral patiency distinction.

[12] By contrast, the "centrist approach" resides more in the analytic philosophy tradition (Gunkel 2007, p. 175).

community of entities considered to have moral status. This particular binary opposition is especially unacceptable to both Gunkel and Levinas (Gunkel 2018b, chap. 3), because for a previously excluded entity—or "Other"— now to be included as deserving moral status, the Other's differences must be reduced to sameness. For Gunkel and Levinas, there are two problems with this. First, those already included (i.e., humans) have the power to admit (or not) the Other into the club of inclusion. Second, when they do admit a previously excluded Other into the club of inclusion, they do so because of the Other's similarity to them based on some property or properties. Thus, one could say that the previously excluded Other is now admitted into the club, but must check its differences at the door—even though those originally inside were not so restricted. As Gunkel states:

> Levinas deliberately interrupts and resists this homology or reductionism, which is, as he argues, an exercise of "appropriation and power" (Levinas 1987, 50). He does not just contest the different universal terms that have been identified and asserted as the common ontological element underlining differences but criticizes the very logic that comprises this *reductio differencia*. "Perceived in this way," Levinas (1969, 43) writes, "philosophy would be engaged in reducing to the same all that is opposed to it as other." (Gunkel 2018b, p. 164)

By "universal terms," Gunkel seems to mean rationality, sentience, life, and any other property proposed by various theories, at one time or another, to be the threshold for having moral status. Ironically, Levinas' Other is strictly human, thereby making even Levinas' theory anthropocentric (Gunkel 2012; 2014; 2018b). Gunkel believes, however, that Levinas' basic insights need not be restricted to the case of human others, but can be applied to the case of non-human others—even the machine other (Gunkel 2014, p. 128).

As with Gunkel's "thinking otherwise" (2007; 2012; 2014; 2018b), the "relational turn" of Coeckelbergh (2009; 2010; 2012) is influenced by continental philosophy. His "relational turn" is an alternative to the "properties-based approach" (Coeckelbergh 2012). He finds the properties-based approach to ascribing moral status lacking, due to intractable epistemological problems.[13] For example, how do we know which property or properties matter for moral status and how do we know why they matter? In addition, because "most properties we hold morally relevant involve a 'mental' aspect" (p. 14), how can we know whether an entity under consideration actually possesses the relevant property or properties?

[13] A similar critique is found in Gunkel (2018b, chap. 3).

In the relational turn of Coeckelbergh (2009; 2010; 2012), ascription of moral consideration relies on primacy of the *relation* between subject and object *within a social context*. For Coeckelbergh, the relation is prior to both subject and object in some sense. Appearance is crucial. The object appears in the consciousness of the subject. This *phenomenological* appearance of the object initiates the relation with the subject by co-determining their interaction. This relation, in a particular social context, gives rise to the subject's moral consideration of the object. Because the relation is the key to determining moral consideration and because this relation is based on appearance, Coeckelbergh's relational turn avoids a certain epistemological problem of the properties-based approach. There is no longer a need to determine whether an object possesses the necessary property or properties required for having moral status. Rather, there need only be a subject's having a phenomenological experience, within a social context, to establish the relation between subject and object.

Because a subject ascribes moral consideration based on relation (and ultimately, based on the object's phenomenal appearance to the subject), the relational turn is an approach Gunkel believes to be consistent with that of Levinas (Gunkel 2014):

> In other words, for Levinas at least, prior determinations of agency and patiency do not first establish the terms and conditions of any and all possible encounters that the self might have with others and with other forms of otherness. It is the other way around. The Other first confronts, calls upon, and interrupts self-involvement and in the process determines the terms and conditions by which and in response to which the standard roles of moral agent and moral patient come to be articulated and assigned. (p. 127)[14]

To summarize briefly, for Gunkel, the "centrist approach" is, in essence, anthropocentric and thereby exclusive. Even when some innovation facilitates greater inclusiveness, it does so by reducing difference to sameness. For Coeckelbergh (as well as for Gunkel), the "properties-based approach" (which essentially is what Gunkel calls the "centrist approach") suffers from several apparently intractable epistemological problems.

2.2 Critique of Challenges to Anthropocentrism

As Rolston (1975) has noted, "an ethic prescribes what ought to be." Undoubtedly, machine intelligence requires a machine ethic. Even if super AI never materializes, even if machines never become conscious, even if we fall short

[14] *Prima facie* this seems similar to a kind of position found in Plato's *Theaetetus* (M. Ananth, personal communication, June 28, 2020).

of human-level machine intelligence/artificial general intelligence, it seems impossible to deny that intelligent machines will become more so. Intelligent machines will increasingly interact with us in moral contexts as well. Such machines will need to be endowed with some functional moral agency (Wallach and Allen 2009). We will continue to ask whether we have obligations to such machines, what these obligations are, and whether such machines would have rights, moral and/or legal. There needs to be a machine ethic.

By design, the other-oriented/relational challenge to anthropocentrism does not qualify as such an ethic. Regarding "thinking otherwise," Gunkel (2012) acknowledges this:

> From one perspective, this outcome cannot help but be perceived as a rather inconclusive kind of ending, one that might not sit well with those who had anticipated and wanted answers or neatly packaged lists of dos and don'ts. … Instead of satisfying this expectation, things have ended otherwise. (p. 211)

Additionally, Coeckelbergh (2012) takes a transcendental approach to his "relational turn," by investigating the "conditions for an entity to appear as having a certain moral status" and "the conditions for moral status ascription/construction" (p. 7). As Gunkel (2013) puts it, "he [Coeckelbergh] is, therefore, not interested in making indubitable claims about the true nature of moral reality but is concerned with tracking and exhibiting the condition for possibility of moral status ascription." Indeed, a Coeckelbergh and Gunkel (2014) collaboration demonstrates how "a relational, Other-oriented approach to moral standing" can be relevant in the area of animal ethics, by illustrating the *conditions* under which some animals are classified as pets (e.g., thus receiving some degree of moral consideration), while other animals are classified as food (e.g., thus receiving a lesser degree of moral consideration). So, the other-oriented/relational challenge to anthropocentrism does not qualify as a machine ethic.

As a sidebar, even though the other-oriented/relational challenge is not intended as an ethic, it does not seem to escape anthropocentrism, either. As mentioned previously, Gunkel (2014) believes that Levinas' basic insights, although definitely anthropocentric, need not be restricted to the case of human others and could be applied in the case of the machine other. However, in further discussing Levinas, Gunkel (2018b) reports that:

> To complicate matters, his "ethics of otherness" has difficulties accommodating and responding to anything other than another human entity. … Furthermore, the subsequent generation of scholars who have followed in Levinas's footsteps have done very little to port his brand of philosophy into technology. (pp.160-161)

“Thinking otherwise” seeks to avoid the moral pitfall of reducing the differences of an Other to the same, as ethical nonanthropocentrism theories supposedly do, when determining that a previously excluded Other does have moral status after all. When the Other is non-human, however, why is this reduction a moral pitfall? Recall that for Levinas, the Other is always another human being (Gunkel 2012; 2014; 2018b). *Prima facie* it seems reasonable that a group of human beings (e.g., classified by gender, race, ethnicity, religion, etc.) traditionally regarded as Other would expect not only moral consideration for its members, but also that this moral consideration include acknowledgment of the differences of that group from the dominant or majority group. When the Other is non-human (e.g., animal or machine) though, what evidence or argument is there that such an entity could or would care if we ignored its differences from us, if this meant gaining moral consideration from us? In short, where is the harm?

As an example, imagine that Regan’s animal rights theory (1985) finally convinces everyone to respect equally the inherent value of any animal that is the experiencing subject of a life. Among other things, humans no longer would eat pigs. Assume that we could devise a way to communicate this happy development to pigs worldwide, with the caveat that we would not be too concerned with their differences from us. Upon learning this, it seems unlikely that pigs could or would care that we are not interested in how they differ from us. They may be perfectly satisfied by the new guarantee of not ending up on our breakfast plates. If the “real” issue, rather, is that *we* should care or that *we* would be harmed by ignoring the differences of the pig, then this pushes us right back into anthropocentrism.

As a second example, consider the intelligent machine case. Suppose human-level intelligence is achieved in some machine. Suppose further that it is endowed with the functional morality described by Wallach and Allen (2009), but lacks consciousness. Arguably, this machine could have “a good of its own” (Scheessele 2018; Basl and Sandler 2013; Kaufmann 1994). Each human (as well as each living thing) also has a “good of its own.” Using this common denominator (i.e., having a good of its own) as a threshold for moral status, the machine could have moral status (Scheessele 2018; Basl and Sandler 2013; Kaufmann 1994). Recall, however, that “thinking otherwise” seeks to avoid reducing difference to sameness in determining moral status of an Other. Surely, our hypothetical machine would differ greatly from humans. One important difference would be its lack of consciousness. To determine that it has moral status, by virtue of having a “good of its own,” would be to ignore this important difference. Because our hypothetical machine is not conscious, though, it cannot care that we have ignored this important difference. If there is some other harm present here, it is not obvious what that might be.

In summary, Gunkel finds morally problematic the disregard by ethical nonanthropocentrism theories for the differences of the Other when reducing these differences to the same for the purpose of extending the Other moral consideration. This concern may be legitimate, but there must be supporting evidence or argument that this is morally problematic in the case when the Other is non-human. Finally, if Gunkel wishes to extend Levinas' philosophy to the "machine question," it seems more development and articulation are necessary.

Regarding ethical nonanthropocentrism theories (e.g., sentiocentrism, biocentrism, ontocentrism), both Gunkel (2018b, chap. 3) and Coeckelbergh (2012, chap. 1) fault such theories for their epistemological problems. These problems include (but are not limited to) knowing which property or properties confer moral status to an entity and knowing whether an entity under consideration actually possesses the property or properties. As Coeckelbergh points out, the latter problem is especially difficult when a morally relevant property has "a 'mental' aspect" (p. 14). As Coeckelbergh and Gunkel (2014) observe, the former problem, that of determining which property or properties confer moral status to an entity, can be inherently anthropocentric (p. 718). Such criticisms of these ethical nonanthropocentrism theories seem largely on the mark. In addition, as already noted in Section 2.1.3, Hargrove (1992) argues that leading ecocentrism theories, those of Callicott (1984) and Rolston (1975), also do not escape anthropocentrism entirely.

3 A Hard Limit on Nonanthropocentrism?

From the previous section, arguably, *ethical* nonanthropocentrism theories may not avoid anthropocentrism entirely. Putting *ethical* anthropocentrism/nonanthropocentrism aside for the moment, recall Thompson's (2017) definition of *conceptual* anthropocentrism: "The idea that human beings can only comprehend the world from a characteristically human perspective—from within a human conceptual framework" (p. 79). This idea implies limitations on human *ethical* nonanthropocentrism. My conjecture is that there is a hard limit on human *ethical* nonanthropocentrism. By "hard limit," I mean that limitations both *exist* and are *exposed*. Limitations on human ethical nonanthropocentrism may exist with respect to moral theorizing about animals, plants, trees, the ecosystem, etc., but these limitations are not exposed (i.e., they are not readily visible) to the extent that we should retreat from appropriate ethical nonanthropocentrism theories, such as those considered in the previous section. Further, I will try to show that the hard limit on human ethical nonanthropocentrism may be justified by the philosophical principle that "ought implies can." Finally, I believe that the potential for super AI has put humanity on a collision course with this hard limit.

3.1 Ought Implies Can

Ought implies can means that *if one ought to do some action* A, *then one can (in the sense of "is able to") do action* A. In Sterba's (2005) explanation of the "ought implies can" principle, "…people are not morally required to do what they lack the power to do or what would involve so great a sacrifice that it is unreasonable to ask, and in cases of severe conflict of interest, unreasonable to require them to abide by" (p. 42). Sterba adds the following stipulation:

> This "ought" implies "can" principle claims that reason and morality must be linked in an appropriate way, especially if we are going to be able to justifiably use blame or coercion to get people to abide by the requirements of morality. (p. 43)

Sterba's account can be split into two separate arguments, **A1** and **A2**. The first argument, **A1 (**let's call it the **standard version of *ought implies can***), may be stated as follows:

P1: If person **X** lacks the power to do action **A**, then person **X** is not morally required to do action **A**.

P2: Person **X** lacks the power to do action **A**.

C: Therefore, person **X** is not morally required to do action **A**.

P1 is just the contrapositive of *ought implies can*. So, if one establishes the truth of **P2**, then **C** follows. If action **A** is to "nonanthropocentrically determine the moral status of an 'other' **Y**," then if one shows that person **X** is incapable of doing this with respect to **Y**, then person **X** is morally off the hook. Establishing the truth of **P2** for action **A**, as it has been defined, would be quite difficult, however. (I will return to this observation in the next section.) Satisfying Sterba's stipulation "that reason and morality must be linked in an appropriate way," on the other hand, immediately follows from the truth of **P1**. If a person **X** lacks the power to do action **A**, then it is unreasonable to morally require that he or she do action **A**.

The second argument based on Sterba's (2005) account, **A2 (**let's call it the **relaxed version of *ought implies can***), may be stated as follows:

P1: If action **A** would involve so great a sacrifice that it is unreasonable to ask, and in cases of severe conflict of interest, unreasonable to require person **X** to abide by, then person **X** is not morally required to do action **A**.

P2: Action **A** would involve so great a sacrifice that it is unreasonable to ask, and in cases of severe conflict of interest, unreasonable to require person **X** to abide by.

C: Therefore, person **X** is not morally required to do action **A**.

Sterba provides a defense of **P1**. By "so great a sacrifice," he does not advocate requiring exclusive focus on "the mere size of the sacrifice" (p. 50). In fact, Sterba does not deny that "…sometimes morality does require great sacrifices from us" (pp. 50-51). But for Sterba there are limits. For example, sacrifices must not be too great in comparison to benefits attained for others via the sacrifices (p. 51). If, in performing an action, the sacrifices of a person (or group) would be too great in comparison to the benefits for another entity (or group of entities), for Sterba, it would be unreasonable to ask the person (or group) to perform the action. It would mean that, in morally requiring performance of action **A**, "reason and morality" *would not be* "linked in an appropriate way." To ground this claim, Sterba argues that, despite their well-known *theoretical* differences, utilitarianism, Kantianism, and Aristotelian virtue ethics, in their "most morally defensible" forms (p. 81), may show little difference in *practical* applications. He appears to be suggesting that no matter the ethical *theory*, there are *practical* limits on the sacrifices we can reasonably ask of others in the "morally most defensible" forms of these various ethical theories (i.e., utilitarianism, Kantianism, and Aristotelian virtue ethics). Yet, Sterba is not wholly disinterested in theory, as Lagerspetz (2007) observes:

> It is not that Sterba sees no use at all for theory. For he thinks morality is rationally justified. It constitutes a reasonable compromise between the egoistic principle "*each person ought to do what best serves his or her overall self-interest"* and pure altruism, as in "*each person ought to do what best serves the overall interest of others"* (pp. 14-15). The rational approach to cases of conflict is to look for non-question-begging solutions, which means that neither egoistic nor altruistic motives are ruled out in advance. High-ranking altruistic reasons should have priority over low-ranking self-interested reasons and vice versa (p. 22). … Similarly, in conflicts between anthropocentric and non-anthropocentric reasons, one needs to find non-question-begging compromises (p. 60). (pp. 189-190)[15]

[15] Still, Lagerspetz (2007) believes that Sterba mistakenly "does not consider that there may be uses for ethical theory other than just their narrowly normative and practical employments" (p. 189). Further, Lagerspetz states that "…Sterba's argument rests on the idea that the only interesting thing about the moral philosophies of Kant, Mill, or Aristotle is, as it were, light theory—to be used for a practical arbitration of whatever issues are being debated among academics or in the media" (p. 189).

So, when Sterba claims that the sacrifices made by one entity (or group of entities) must not be too great in comparison to the benefits attained by another entity (or group of entities), he is not making a simple utilitarian argument measuring total suffering (or pleasure). Rather, he insists on a "reasonable compromise" between a conflicting pair of principles such as egoism/altruism or anthropocentrism/nonanthropocentrism. In summary, this seems to be a plausible defense of **P1** of the relaxed version of *ought implies can*.

To defend the truth of **P2** (of the relaxed version of *ought implies can*) in the case where intelligent machines represent the "other," **Y**, I propose a thought experiment and then compare/contrast this case with the prospect of super artificial intelligence.

3.2 A Thought Experiment

In the novel *Planet of the Apes* (Boulle 1963), three French astronauts discover an Earth-like planet, *Soror*. The only apparent difference between Earth and Soror is that on the latter planet, the great apes are in charge, and human beings are feral. Humans in the wild are hunted by gorillas. Captive humans are experimented on or displayed in zoos. One astronaut is promptly killed in one of the hunts; another is captured and reverts to a primitive state, like that of the men on Soror. The third astronaut, Ulysse, is also captured, but keeps his wits about him, convincing the sympathetic chimp-scientist Zira and her chimp-scientist fiancé Cornelius that his intelligence is comparable to theirs. Ulysse mates with a native Soror human, the feral Nova, who bears a son, Sirius. Before Zira, Cornelius, and friends help the human trio regain the confiscated spaceship in order to escape to Earth, Ulysse and Cornelius learn the truth of what had happened on Soror.

More than 10,000 years before, human beings on Soror were in charge and were like Earth humans. The great apes were feral then. For unknown reasons, the population of the great apes increased. Humans domesticated the apes, many of whom learned to speak and to perform human-level tasks, some even becoming household servants. Increasingly, this new "ape-technology" caused humans to become lazy. Increasingly, the apes refused to take orders from their human masters. This eventually led to an ape-takeover of Soror. The lazy humans fled the cities to remote regions and became animal-like. The great apes mimicked their former masters, taking their places in Civilization 2.0.

Ulysse, Nova, and Sirius escape Soror and return to Earth. When the trio land in Paris, they are greeted by military apes. So, they re-board their spacecraft and head back into space.

Now, imagine an alternate ending to this story. Assume Ulysse and family return to Earth before the ape-takeover gains traction. Further, assume Ulysse has access to a taste-free, scent-free potion that stunts ambition in apes, but stimulates it in humans.[16] He hatches a plan to produce enough potion, of sufficient potency, to introduce it widely into Earth's water supplies. *Prima facie* it seems reasonable for Ulysse to execute his plan to rescue humanity. Suppose, however, that a utilitarian calculation, which takes into account the suffering and pleasure of humans and apes, indicates that Ulysse *ought not* execute his plan.[17] Then what?

In support of **P2** (of the relaxed version of *ought implies can*), morally requiring Ulysse to forgo execution of his plan seems, intuitively at least, to involve so great a sacrifice that it would be unreasonable to ask of him. Especially, the threshold of a "severe conflict of interest" seems to be met such that it would be unreasonable to require Ulysse to abandon his plan. After all, Ulysse is a member of *Homo sapiens*, which stands to suffer greatly if Ulysse does not intervene. *Homo sapiens* has evolved under the same laws of natural selection as other species, laws that act to promote growth, reproduction, and survival of species. *Prima facie* it seems reasonable and natural for Ulysse to ignore the result of a nonanthropocentric utilitarian calculation by going through with his plan to save humanity at the expense of increased welfare for Earth apes. Although the thought experiment was deliberately set up to favor the (non)action where apes come out ahead, provided one subscribes to a strictly utilitarian view, should one really blame Ulysse if he were to execute his plan to keep humans on top? Clearly, on biological grounds and the corresponding conflict of interest, it is not unreasonable to say "no."

[16] In the novel, Boulle (1963) describes the ape-takeover (at least partially) in terms of human abdication. For example, one woman's experience with her once-loyal, long-time gorilla servant goes like this: "I was too frightened. I could not go on living like this. I preferred to hand the place over to my gorilla. I left my own house."

[17] The novel suggests that the ape population at the time of takeover could be approximately equal to the human population. It also suggests that the ape-takeover would be quick (e.g., on Soror, it seemed to take no more than a few years, if that long) and not too bloody (e.g., see previous footnote). However, in this hypothetical scenario, a utilitarian calculation might still be difficult. For example, if Ulysse executes his plan, humans would be in control and would still have talking, domesticated ape-servants. These would simply be less ambitious and more subservient. If Ulysse does not execute his plan, apes would be in control, but may not (initially) have talking, domesticated, subservient human-servants (e.g., due to the flight of humans out of populated areas). Similarly, if Ulysse does not execute his plan, many humans suddenly would be faced with the hardship of primitive conditions. However, some could also regain their previous vigor and initiative as they acclimate to these conditions. On the other hand, if Ulysse executes his plan, it is possible that now-subservient, talking, domesticated ape-servants would be trapped in a limbo-state between their previous freedom as wild animals (i.e., prior to their domestication) and a now-denied opportunity to develop full autonomy. No doubt there are other issues that would make utilitarian calculations difficult in this hypothetical scenario. For the purposes of my thought experiment, though, it does not seem any more unreasonable to suppose that a utilitarian calculation would not favor Ulysse's plan versus favor it (or be indifferent to a choice between the two).

Of course, this “Planet of the Apes Scenario” (POTAS) is virtually impossible, so humans have not had the opportunity to glimpse the existence of a hard limit. As a result, we may be currently free to (and arguably should) rant against anthropocentrism in an attempt to eliminate it from moral decision-making. However, the prospect of super AI seems to parallel POTAS. We may have the opportunity to confirm (or disconfirm) the existence of a hard limit after all.

3.3 Super AI

Bostrom (2014/2016) combined the results of several surveys of experts and found that 90% believe that human-level machine intelligence (HLMI) is achievable by 2075; 75% believe, given achievement of HLMI, that superintelligence will follow within thirty years.[18] Although Bostrom does not argue that we are on the threshold of super AI or that we even could predict such an event with precision, he seems to believe a breakthrough is “somewhat likely” and “sometime in this century” (Preface, pp. v-vi). Unlike POTAS, then, the prospect of super AI seems plausible.

Bostrom (2014/2016) suggests the possibility of an “intelligence explosion,” where modest AI improves to the point of super AI. Human-level machine intelligence (HLMI) would occur between these two points. He observes that, once HLMI occurs, the “takeoff” leading to super AI could be slow (on the order of decades or centuries), moderate (on the order of months or years), or fast (on the order of minutes, hours, or days). He questions whether one superintelligence would emerge first or whether multiple teams would produce a number of distinct superintelligences at roughly the same time. Bostrom asks, “will the frontrunner get a decisive strategic advantage?” (p. 96), where “decisive strategic advantage” is defined as “strategic superiority (by technology or other means) sufficient to enable an agent to achieve complete world dominance” (p. 407). Further, Bostrom outlines “an AI takeover scenario” (chap. 6) for the reader. He does not regurgitate and synthesize bits and pieces of “Terminator-style” scenarios, but rather thoughtfully paints a plausible picture of what could happen.

[18] Bostrom (2014/2016) cautions that small sample sizes and other methodological issues do not permit drawing “strong conclusions” from these results (p. 25). Human-level intelligence (in a machine) is roughly equivalent to what some refer to as “strong AI” or “artificial general intelligence” (AGI). See Bostrom (2014/2016, p. 22) for a brief discussion. Note that in this context, it does not seem to be implied that a strong AI (or AGI) must be conscious. There also seems to be no distinction made between a strong AI that thinks and is intelligent versus one that merely simulates thinking and intelligence. A machine with human-level intelligence would exhibit at least as much intelligence as a typical human being in a broad number of domains. Concerning “superintelligence,” Bostrom considers several forms which this could take, including super artificial intelligence (AI), cognitively enhanced humans, sophisticated brain-computer interfaces, etc. For this essay, I emphasize his discussion of super AI. “Intelligence” of a super AI would greatly surpass human intelligence in most domains.

Although many have questioned whether a silicon-based machine intelligence could be conscious, sentient, able to feel, etc., and if so, how humans might know this is the case, Bostrom takes a different direction. He considers the *motivations* of a super AI, by asking, “But what will its goals be?” (Bostrom 2014/2016, p. 127) Referring to the human tendency to anthropomorphize fictional aliens as well as AI, he elaborates:

> An artificial intelligence can be far less human-like in its motivations than a green scaly space alien. The extraterrestrial (let us assume) is a biological creature that has arisen through an evolutionary process and can therefore be expected to have the kinds of motivation typical of evolved creatures. It would not be hugely surprising, for example, to find that some random intelligent alien would have motives related to one or more items like food, air, temperature, energy expenditure, occurrence or threat of bodily injury, disease, predation, sex, or progeny. …
>
> An AI, by contrast, need not care intrinsically about any of those things. There is nothing paradoxical about an AI whose sole final goal is to count the grains of sand on Boracay, or to calculate the decimal expansion of pi, or to maximize the total number of paperclips that will exist in its future light cone. In fact, it would be *easier* to create an AI with simple goals like these than to build one that had a human-like set of values and dispositions. (pp. 128-129)[19]

The underlying worry here seems to be that a super AI, with a decisive strategic advantage, armed with *whatever* goal, could be motivated to pursue its goal relentlessly. Thus, in contrast to POTAS, the consequences of super AI appear to be less predictable. The possibility of *Homo sapiens* becoming extinct due to super AI could be greater than with a POTAS. Intelligent apes, as with the “green scaly space alien” in Bostrom’s example, would have “arisen through an evolutionary process and can therefore be expected to have the kinds of motivation typical of evolved creatures.” Like many humans, some intelligent apes of a POTAS would probably have a tenderness toward animals of other species, rendering total extinction of *Homo sapiens* unlikely. A powerful super AI relentlessly pursuing some goal, on the other hand, might act as if oblivious to such concerns. This seems especially true if a

[19] The first day of the inaugural AAAI/ACM conference on Artificial Intelligence, Ethics, and Society, held in New Orleans, LA, USA, Feb. 1-3, 2018, focused largely on the “value alignment” problem. The systems under consideration were domain-specific decision-making systems trained on data from past human transactions. Machine learning techniques used to create such systems have shown a tendency to incorporate human bias (racial, gender, etc.) gleaned from the training data into the final decision-making systems intended to be deployed for use by society. “Value alignment” research aims at remedying this problem.

dominant super AI is either not conscious[20] or has an emergent (or perhaps artificial) consciousness that differs greatly from that of biological creatures.

Bostrom (2014/2016, chap. 9) describes two major classes of approaches for preventing the existential threat of a super AI with decisive strategic advantage and perhaps difficult-to-understand motivations. The first class of approaches to solving this so-called “control problem” focuses on development of super AI containment techniques. For instance, “boxing methods” aim at constraining a super AI’s ability to interact with the world external to it (whether physical or informational). “Incentive methods” aim at putting a super AI into an environment with appropriate incentives for it to play nicely with others. “Stunting” could “limit the system’s intellectual faculties or its access to information.” “Tripwires” could detect when a system has exceeded certain thresholds such that it possibly could be shut down. The second class of approaches targets super AI motivations. For example, “direct specification” could “explicitly define a set of rules or values that will cause even a free-roaming superintelligent AI to act safely and beneficially.” Clearly, such attempts to solve the control problem are anthropocentric, given their emphasis on preventing a super AI, particularly one with decisive strategic advantage, from posing an existential threat to humanity.

3.4 Analysis

Not only does super AI appear to be more plausible than the Planet of the Apes Scenario, it also is potentially more unpredictable than POTAS. As with Bostrom’s (2014/2016) green alien example, we would have some insight into the motivations and goals of apes that we might lack in the case of a super AI. Further, as with Zira and Cornelius, some apes could be sympathetic, preventing complete human extinction and perhaps providing some comfort to pockets of humans here and there. In the case of super AI, however, given its superior intelligence and potentially unpredictable motivations and goals, it could pose an existential threat to humans. Asking humans *not* to be anthropocentric in moral deliberations anticipating super AI would seem to meet the threshold of “so great a sacrifice” (Sterba 2005). Even if super AIs could feel, and thus suffer or have pleasure in some meaningful sense, there likely would be fewer super AIs than humans. Thus, any net benefits to one super AI (or to a few super AIs) could pale in comparison to the overwhelming sacrifice of humans. Adding to this sacrifice would be the potential loss of an entire species, *Homo sapiens*, and everything entailed by that. This would justify **P2** of **A2**, showing that

[20] See Torrance (2008) for a related observation involving artificial agents and sentience.

humans should not be obligated to be completely nonanthropocentric in determining the moral status of a super AI. This is the hard limit on human nonanthropocentrism.

The current human response, as manifested by various approaches to solving the super AI control problem, is clearly anthropocentric. This response might show merely what currently *is* the case, rather than what *ought* to be the case. On the other hand, it could suggest that a hard limit on human nonanthropocentrism is already manifesting itself. When we contemplate the moral status of a future super AI, given even a small chance of its creation and a small chance that it might pose an existential threat to humans, it seems perfectly reasonable that at least some anthropocentrism would influence our deliberations. Acting nonanthropocentrically may seem to be what we *ought* to do, but this may be unreasonable if it requires too "great a sacrifice," as Sterba (2005, p. 42) might put it.

4 Potential Objections

There are at least three possible objections to an argument for a hard limit on human nonanthropocentrism. The first objection may concede that there is a hard limit on human nonanthropocentrism, but question whether truth of this conjecture would negatively impact arguments for animal rights, biocentrism, etc. The second objection may be concerned with the implications, if the conjecture of a hard limit is shown to be false. The third objection may contest my use of what I have called the "relaxed version" of Sterba's account (2005) of "ought implies can" in order to argue that there is a hard limit on human nonanthropocentrism.

4.1 What If the Conjecture of a Hard Limit Proves to Be True? Would This Negatively Impact Animal Rights, Biocentrism, etc.?

If there exists a hard limit on human nonanthropocentrism, one might object that this would lead to less concern with anthropocentric tendencies when dealing with animals, plants, trees, the land, etc. Further, the objection might go, if there is a hard limit on human nonanthropocentrism, why bother at all? The result of such an attitude could slow or even halt progress in extending moral consideration to non-human others. This is a slippery slope worry.

This does not have to be the case, however. As argued earlier, the prospect of super AI poses a novel challenge for humans. This challenge entails the possibility, however small, of the domination, and perhaps extinction, of *Homo sapiens* by one or more super AIs. Certain intelligent technology with some degree of autonomy could bring us to a hard limit in ways that animals, plants, trees, the land, etc., never could. Thus, it does not necessarily follow that we should abandon attempts to overcome anthropocentrism when dealing with these non-human, non-machine others.

One way to salvage the objection may be to claim that certain species of animals must have posed similar challenges to our ancestors as super AI may pose for us. “Sure,” this criticism continues, “now we have dominion over other animals, and the probability of an event such as POTAS is negligible. But our ancestors surely did not have the control over other animals that we enjoy.” Bostrom’s (2014/2016) example of the green alien, who has evolved via a similar process of evolution as *Homo sapiens*, can address this criticism though. We would have some immediate insights into probable motivations and goals of such a creature. Presumably, the same was true for our ancestors. Even though they did not know about evolution, perhaps observation, mixed with some intuition, informed them about other animals’ ability to suffer, efforts to survive—in short, about their motivations and goals. Plausibly, our human ancestors may have recognized these motivations and goals as similar to their own. Thus, human ancestors would not have faced a challenge comparable to that of super AI, the motivations and goals of which may not be obvious to humans. In sum, the slope is not as slippery as this objection might suggest.

To bolster this conclusion further, occasional conflict between modern humans and non-human animals can offer a glimpse into how human ancestors might have coped. Along the Chobe River, separating Namibia and Botswana in south-central Africa, deadly crocodile attacks on humans dramatically increased in the first decade of the 21st century (Cole 2014). These crocodiles, one of the most intelligent reptile species, ate human children and attacked livestock in human settlements adjacent to the Chobe. Understandably, villagers were upset at their inability to control these large, deadly, intelligent animals that threatened them and their livelihoods. One father, whose small son had been devoured by a crocodile, relayed to the documentary-producers that he wanted to kill crocodiles. This sentiment, understandably, was not uncommon. Three scientists visited the settlement and developed a solution that satisfied the villagers, keeping them safe while also protecting crocodiles from their anger. The solution was based on early 20th century psychology discoveries: classical conditioning and social learning. They connected bait to a bell that rang just before a crocodile could reach the bait. Coupled with the bell ring was an electric shock to the crocodile’s snout. Crocodiles learned quickly to retreat from an area at the sound of the bell’s ring. Because this species of crocodile has a dominant male leader keeping watch over its group’s territory, the scientists would identify the dominant croc and lure it to the trap with the bait, bell, and shock. Non-dominant crocodiles in a territory quickly learned from seeing their leader shocked, such that they, too, began to retreat at the sound of the bell. After conditioning the crocs in a territory, just sounding the bell was sufficient to keep crocs away from the banks of the Chobe. Periodically, the scientists repeated the conditioning as a sort of “booster shot” for a territory. Of course, human ancestors would not have had formal knowledge of classical conditioning and social learning techniques to employ against predatory or nuisance animals. However, the key insights for the team of scientists were that the crocs are attracted by food (the bait) and that they likely could feel pain and would respond aversively to it. It seems reasonable that human ancestors, over time, also would have detected these reactions and behaviors of non-human animals and would have exploited this information to control non-human animals as an alternative to killing them. The scientists were simply more effective and efficient because they could draw on well-established psychological knowledge and techniques. As this scenario illustrates, human ancestors likely did not face a challenge quite like that of super AI. If so, human ancestors would not have faced the hard limit of human nonanthropocentrism.

4.2 What If There is No Hard Limit on Human Nonanthropocentrism?

This could happen in two ways. In the first case, assume super AI is realized, but my conjecture of a hard limit in the face of super AI is wrong. In the second case, assume super AI is never realized, such that a hard limit would not materialize. One may then object that if a hard limit is assumed, this assumption might unnecessarily promote anthropocentrism in determining the moral status of *all* intelligent machines, whether super AI or not. If humans actually were to have moral obligations to intelligent machines, any promotion of anthropocentrism could lead to wronging these machines.

Although this is a possibility, several mitigating factors work in favor of intelligent machines. First, several views already pave the way for moral consideration of intelligent machines (Gunkel 2012; Scheessele 2018; Tavani 2018; Gerdes 2015; Coeckelbergh 2010). Floridi (2008) argues that anything that exists (including intelligent machines) “can place moral claims on the interacting agent and ought to contribute to the constraint and guidance of his ethical decisions and behavior” (p. 48). Kaufmann (1994) argues that machines (not just intelligent machines) have interests. According to Kaufman, because benefits or harms only matter to entities with interests, it is just such entities that are candidates for moral status. Basl and Sandler (2013) argue that artifacts (which would include intelligent machines) can have “a good of their own,” making them eligible for moral status. They also point out how synthetic biology has blurred the distinction between artifacts and naturally-occurring organisms. Thus, several lines of research have already laid a foundation for moral consideration of intelligent machines. This challenges anthropocentrism. So, these research efforts tell against the cogency of this second objection.

Further, human capacity for anthropomorphizing[21] objects is well-known. This could work to the advantage of many "species" of intelligent machines. Darling (2017) even suggests "anthropomorphic framing" of select types of robots when this framing "directly supports the main function of the robot" (p. 183). Framing a certain type of robot in such a way as to promote our natural tendency to anthropomorphize it could lead to extending the robot moral consideration. This may still be a form of anthropocentrism, but moral consideration would be extended beyond humans nonetheless.

One may protest that humans often anthropomorphize animals of various species, yet, overall, animals do not enjoy the moral status that humans do. However, many categories of intelligent machines will continue to have the ability to communicate directly with humans. This gives them an advantage over animals, plants, trees, etc. As Gunkel (2018b) explains, there are two main classes of rights theories: "interest" theories and "will" theories:

> Interest theories connect rights to matters of welfare. … "Will" theorists, by contrast, require that the subject of a right possess the authority and/or capacity to assert the privilege, claim, power, or immunity. (p. 31)

Gunkel's analysis suggests that "interest" theories could be useful in justifying rights for animals and other moral patients. By contrast and by virtue of language backed by rationality, some categories of intelligent machines could demand rights for themselves and other species of intelligent machines. "Will" theories would be more relevant for such cases. In short, it may be possible for rational, language-capable intelligent machines (and their supporters) to leverage existing "will" theories in order to gain rights for themselves and other species of intelligent machines.

To summarize: The second objection asks what might happen if we assume a hard limit that ultimately does not exist (or does not materialize). Would this needlessly promote anthropocentrism, thus leading to the harm of intelligent machines? Despite this possibility, several factors, including various lines of research suggesting moral consideration for intelligent machines, potential human tendency to anthropomorphize some intelligent machines, plus anticipated ability of some intelligent machines to communicate directly and rationally with humans, even to argue directly for their own rights, weaken the force of this objection.

4.3 Proof of a Hard Limit Must be Supported By the "Standard Version" of "Ought Implies Can"

[21] Anthropomorphization is the human tendency to attribute human characteristics to non-humans. Although this tendency is anthropocentric, to the extent that it causes humans to extend moral consideration to non-humans, this tendency is nonanthropocentric in its effect.

In the previous section, I supported my conjecture of a hard limit on human nonanthropocentrism using what I have called the "relaxed version" of Sterba's account (2005) of "ought implies can." This version of Sterba's account is not without critics, however. For example, one criticism by Lagerspetz (2007) has to do with how a conflicting pair of principles, such as egoism/altruism, is selected in Sterba's account. (Recall that Sterba's conflict resolution would involve a compromise between such a conflicting pair of principles.) Critics of Sterba's account of "ought implies can" may insist that any argument for a hard limit rest on the "standard version," not the "relaxed version," of the "ought implies can" principle. As observed in the previous section, this would require showing **P2** of argument **A1**: Person **X** lacks the power to do action **A**, where action **A** is to "nonanthropocentrically determine the moral status of an 'other' **Y**." It would be difficult to prove the truth of **P2** in argument **A1**, though. Nevertheless, one approach may be to find evolutionary support.

If "lacks the power" in **P2** of argument **A1** (the "standard version" of *ought implies can*) were to have a biological basis, and if an evolutionary explanation were available for this biological basis, it would be strong support for **P2** of **A1**. The implication is that one could reasonably claim a hard limit on human nonanthropocentrism that is justified by the "standard version" of *ought implies can*. Natural selection promotes traits that facilitate the survival, growth, and reproduction of a species' members. Numerous species have appearances that camouflage them or hard-wired behaviors that help them avoid predators. Unfortunately, it is not obvious that some biological trait exists to ground a hard limit on human nonanthropocentrism. Fortunately, in addition to special-purpose traits, *Homo sapiens* boasts perhaps the most impressive evolutionary product of all, human cognition. Human cognition acts as a general-purpose facilitator of human survival, growth, and reproduction by enabling us to reason, solve problems, plan, generate counterfactuals, engage in language, etc. So, when faced with the prospect of a future super AI, particularly one that could gain a decisive strategic advantage, intuition suggests that we would use our general-purpose counterfactual-generating, problem-solving human cognition to prevent whatever threat such a super AI might pose to our existence.

Is there some function of a general-purpose mind that would prevent humans from being completely nonanthropocentric in the face of super AI? Even if there is such a cognitive function, Lewontin (1998) cautions that, while human cognition no doubt is a product of evolution, giving an evolutionary account of a specific human cognitive function usually amounts to little more than "plausible storytelling" (p. 129). Similarly, Ananth (2018) analyzes the "strong" version of evolutionary psychology (SEP), "the ambitious research crusade that not only employs evolutionary theory, but does so in a way that is committed to a unique set of principles designed to reveal the nature of human mind and behavior" (p. 255). Ananth concludes that analysis "has revealed that SEP's ambitions have far exceeded the evidence proffered" (p. 273). Thus, even if we could identify some cognitive function acting to prevent us from being completely nonanthropocentric in the face of a potentially threatening Other, such as a super AI, Ananth and Lewontin caution that it would be difficult to establish rigorously an evolutionary biological trail for this cognitive function. Establishing such a biological trail would be especially difficult, given cultural as well as environmental influences along this trail. Further, if it were possible to establish rigorously the evolution of such a cognitive function, it may not mean that this function *determines* our actual behavior. Instead, the function may serve more as a *disposition* to behave a certain way. Summing up, it may not be possible to establish rigorously a cognitive function that would make it *impossible* for humans to completely forgo anthropocentrism in consideration of the moral status of a non-human Other, such as super AI. That is, it may not be possible to *prove* **P2** of **A1** (from the preceding section) that "Person **X** lacks the power to do action **A**," where action **A** is to "nonanthropocentrically determine the moral status of an 'other' **Y**." The best that one may be able to do is to construct a "plausible story" to suggest the truth of **P2**. The quest for this plausible story is the topic of a current project.

Even if one could construct such a "plausible story," the "standard version" of *ought implies can* requires understanding the usage of 'can.' Ananth (M. Ananth, personal communication, June 28, 2020) explains: "In the ought implies can discussion, a lot turns on what is meant by 'can'. If 'can' means "is able to based on resources available," then there might be a problem. Imagine that I intentionally incur a very large debt, knowing full well that I could never re-pay it. On this version of 'can', I am off the hook to pay this debt because I do not have the resources to pay this debt." This concern has implications for the case of super AI. Suppose one discovers a biological trait (or cognitive function) that plausibly would limit our ability to be nonanthropocentric in determining the moral status of a super AI. Given such knowledge, it could be morally problematic to pursue development of super AI. That is, if we were aware of a likely hard limit on human nonanthropocentrism in the face of super AI, and if anthropocentrism to any degree is problematic[22], then perhaps we ought to abandon pursuit of super AI.[23] This seems even more advisable if a super AI could suffer.

5 Conclusion

My conjecture is that there is a hard limit on human nonanthropocentrism. This should not disturb the animal rights debate or environmental debates. My conjecture is that we face the hard limit on human nonanthropocentrism given the prospect of super AI, and perhaps even human-level machine intelligence. So-called "control problem" research is anthropocentric. This research likely is not merely a description of what *is* the case, but serves as an example of what *ought to be* the case if there is a non-negligible chance that super AI would pose an existential threat to humanity. In any case, if anthropocentrism would wrong or harm a super artificial intelligence and if the hard limit proposed here exists, then we ought not pursue super AI and we ought to act to prevent its emergence.

[22] Anthropocentrism constituting the "weak anthropocentric intrinsic value" of Hargrove (1992) may be an exception.

[23] This seems in the same spirit as Bryson (2010), who argues "that it would… be wrong to build robots we owe personhood to."

References

Ananth, M. (2018). *Bringing biology to life: An introduction to the philosophy of biology*. Tonawanda, NY: Broadview Press.

Anderson, M., & Anderson, S.L. (2007). Machine Ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4). 15-26.

Basl, J., & Sandler, R. (2013). Three puzzles regarding the moral status of synthetic organisms. In G. E. Kaebnick & T. H. Murray (Eds.), *Synthetic biology and morality: Artificial life and the bounds of nature* (pp. 89-106). Cambridge, MA: MIT Press.

Bostrom, N. (2014/2016). *Superintelligence: Paths, dangers, strategies*. New York: Oxford University Press.

Boulle, P. (1963). *Planet of the apes.* New York: Random House Publishing Group. (Translated from French to English by Xan Fielding.)

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: key social, psychological, ethical and design issues* (pp. 63-74). (Natural Language Processing; Vol. 8). John Benjamins Publishing Company.

Callicott, J. B. (1984). Non-anthropocentric value theory and environmental ethics. *American Philosophical Quarterly, 21*(4). 299-309.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society, 24*. 181-189.

Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology, 12*. 235-241.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. New York: Palgrave Macmillan.

Coeckelbergh, M., & Gunkel, D.J. (2014). Facing animals: A relational, other-oriented approach to moral standing. *Journal of Agricultural and Environmental Ethics, 27*. 715-733.

Coeckelbergh, M., & Gunkel, D.J. (2016). Response to "The problem of the question about animal ethics" by Michal Piekarski. *Journal of Agricultural and Environmental Ethics, 29*. 717-721.

Cole, S. (Producer). (2014). *Operation maneater. Episode 1: Crocodile*. Windfall Films.

Darling, K. (2017). “Who's Johnny?” Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. In P. Lin, R. Jenkins, K. Abney (Eds.), *Robot Ethics 2.0: From autonomous cars to artificial intelligence* (Ch. 12). New York: Oxford University Press.

DeGrazia, D. (2002). *Animal rights: A very short introduction.* New York: Oxford University Press.

DesJardins, J.R. (2015). Biocentrism. In “The Editors of Encyclopaedia Britannica” (Eds.) *Encyclopaedia Britannica*. Chicago: Encyclopaedia Britannica, Inc. URL: https://www.britannica.com/topic/biocentrism. (Retrieved on Jan. 16th, 2020.)

Faria, C., & Paez, E. (2014). Anthropocentrism and speciesism: Conceptual and normative issues. *Revista de Bioetica y Derecho, 32*. 82-90.

Floridi, L. (2008). Information ethics: Its nature and scope. In J. Van Den Hoven & J. Weckert (Eds.), *Information technology and moral philosophy* (Ch. 3). New York: Cambridge University Press.

Floridi, L., & Sanders, J.W. (2004). On the morality of artificial agents. *Minds and Machine*, *14*. 349-379.

Floridi, L., & Taddeo, M. (2018). Don’t grant robots legal personhood. *Nature*, *557*. 309.

Gerdes, A. (2015). The issue of moral consideration in robot ethics. *ACM SIGCAS Computers and Society*, *45*(3). 274-279.

Gunkel, D. (2007). Thinking otherwise: Ethics, technology and other subjects. *Ethics and Information Technology, 9*. 165-177.

Gunkel, D. (2012). *The machine question: Critical perspectives on AI, robots, and ethics.* Cambridge, MA: MIT Press.

Gunkel, D. (2013). Review of Mark Coeckelbergh's *Growing Moral Relations* (Palgrave, 2012). *Ethics and Information Technology*, *15*(3). 239-241.

Gunkel, D. (2014). A vindication of the rights of machines. *Philosophy & Technology*, *27*. 113-132.

Gunkel, D. (2018a). The other question: Can and should robots have rights? *Ethics and Information Technology, 20*(2). 87-99.

Gunkel, D. (2018b). *Robot rights*. Cambridge, MA: MIT Press.

Hargrove, G. (1992). Weak anthropocentric intrinsic value theory. *The Monist*, *75*. 183-207.

Kaufman, F. (1994). Machines, sentience, and the scope of morality. *Environmental Ethics*, *16*(1). 57-70.

Lagerspetz, O. (2007). [Review of the book *The triumph of practice over theory in ethics*, by J. Sterba]. *Philosophical Investigations, 30*(2). 188-191.

Leopold, A. (1949/1977/2010). A Sand County almanac: The land ethic. In G. Marino (Ed.) *Ethics: The essential writings*. (pp. 487-505). New York: Modern Library. (Reprinted from *A Sand County almanac*, pp. 201-206, 1949/1977, New York: Oxford University Press)

Lewontin, R. C. (1998). The evolution of cognition: Questions we will never answer. In D. Scarborough & S. Sternberg (Eds.) *Methods, models, and conceptual issues: Vol. 4. An invitation to cognitive science*. (pp. 106–132). Cambridge, MA: The MIT Press.

Prodhan, G. (2016). Europe's robots to become 'electronic persons' under draft plan. Reuters.com. (*Science News*: June 21, 2016). https://www.reuters.com/article/us-europe-robotics-lawmaking/europes-robots-to-become-electronic-persons-under-draft-plan-idUSKCN0Z72AY Retrieved on March 8th, 2018.

Rae, G. (2016). Anthropocentrism. In H. ten Have (Ed.) *Encyclopedia of global bioethics*. Cham, Switzerland: Springer.

Regan, T. (1985). The case for animal rights. In P. Singer (Ed.) *In defense of animals*. (pp. 13-26). New York: Basil Blackwell.

Rolston III, H. (1975). Is there an ecological ethic? *Ethics, 85*(2). 93-109.

Scheessele, M.R. (2018). A framework for grounding the moral status of intelligent machines. In *Proceedings of 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*, February 2-3, 2018, New Orleans, LA, USA. New York: ACM.

Singer, P. (1974). All animals are equal. *Philosophic Exchange, 5*(1). 103-116.

Singer, P. (1985). Ethics and the new animal liberation movement. In P. Singer (Ed.) *In defense of animals*. (pp. 1-10, 209-211). New York: Basil Blackwell.

Sterba, J.P. (2005). *The triumph of practice over theory in ethics.* New York: Oxford University Press.

Tavani, H.T. (2018). Can social robots qualify for moral consideration? Reframing the question about robot rights. *Information, 9*(4). doi:10.3390/info9040073.

Taylor, P.W. (1981/2010). The ethics of respect for nature. In L. Vaughn (Ed.) *Doing ethics: Moral reasoning and contemporary issues, second edition*. (pp. 512-526). New York: W.W. Norton & Company. (Reprinted from *Environmental Ethics, 3(3)*, pp. 197-218 (edited), 1981)

Thompson, A. (2017). Anthropocentrism: Humanity as peril and promise. In S.M. Gardiner and A. Thompson (Eds.) *The Oxford handbook of environmental ethics*. (pp. 77-90). New York: Oxford University Press.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & Society, 22*. 495-521.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.