Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines[1][2]

Introduction

On any reasonable account of who or what is a moral patient – i.e., who or what has a welfare that we must consider in our moral deliberations – once we create an artificial consciousness with capacities like our own that experiences the world much like we do – we must recognize that consciousness as a moral patient; it will be due consideration for its own sake in virtue of the fact that it has interests in the kinds of things we do and that we take to be morally relevant in our kind and because there will be little or no reason to discount those interests. [3] Indeed, it seems plausible that insofar as artificial consciousness approximates our own mental life, it will be due equal consideration whether that is understood in consequentialist, deontological, or other ways.[4]

However, we are a long way from creating an artificial consciousness that is anything like our own, or, for that matter, perhaps from creating artificial consciousness unlike our own. Yet, as we create more and more self-directed, self-maintaining machines and attempt to create artificial consciousness, we must think carefully about which properties of a machine would confer interests or moral patiency. To fail to ask whether these entities have interests and whether they are due consideration may lead to inappropriate conduct on our part. After all, it is not only beings with a consciousness like ours that are moral patients; non-human

[1] Acknowledgments

[2] Some of the ideas and language, particularly definitions, also appear in REMOVED TO PROTECT ANONYMITY.

[3] I talk in terms of consciousness rather than intelligence to avoid taking a stand on the relationship between the two. I assume instead that it is possible for a machine to be intelligent without it being conscious.

[4] What exactly that means for the treatment of such beings will be a function of their nature and our relationships to them, just as equal consideration of humans is sensitive to these factors.

animals are moral patients, and we owe it to them to take their interests into account in our moral deliberations. This is so even though their mental life may be radically different than our own. This forces us to determine (i) under what conditions machines have a welfare and (ii) whether and how to take that welfare into account.

As difficult as it might be to build an artificial consciousness, we also face great difficulties in answering the philosophical questions raised by the creation of such an entity. Above I raised both a metaphysical question (what properties of machines constitute their having a welfare or interests) and a normative question (how are we to take the welfare of artificial consciousness into account). There are also serious epistemic questions we must answer. How are we to know when a machine is conscious? What are the sources of evidence for consciousness? In what follows, I take up these challenges with a primary focus on the metaphysical and epistemological questions. In exploring these challenges, I argue that despite the philosophical difficulties we face, in fact, perhaps because of them, we are licensed in behaving as if current machines are not moral patients.

In Section I, I explain the concept of moral status and its relationship to the concept of welfare and interests. In Section II, I take up the metaphysical, epistemic, and normative issues concerning the moral status of artificial consciousness. I argue that in order for an entity to have what I will call *psychological interests* it must have the capacity for attitudes.[5] Entities might be conscious even if they lack this capacity, but they will not have any interests in virtue of that consciousness. In section III, I argue that current machines are *mere machines*, machines

[5] It seems plausible that both the capacity for desires and the capacity for preferences requires or is partly constituted by the capacity for attitudes. If this is false, then attitudes can be understood as “attitudes, preferences, or desires” throughout this paper.

that lack the capacity for attitudes. Despite the fact that current machines may have all sorts of capacities, we have no evidence whatsoever that they are conscious, let alone that they have attitudes regarding their conscious states. Even those skeptical of this conclusion, I argue, will have to agree that we have no evidence whatsoever what their attitudes are. Given this, for all practical purposes, any interests that these machines have are irrelevant to our moral deliberations. To deny this is to violate the principle that we must be excused for any violation of obligations that we couldn't possibly know that we have given our capacities and evidence.[6]

I do not wish to argue that an entity has a welfare only in virtue of having psychological interests. Many environmental ethicists have argued that non-sentient organisms have interests in virtue of being goal-directed, or teleologically organized systems. In Section IV I explore the possibility that machines have what I'll call *teleo interests* in virtue of their being teleologically organized. In Section V, I will argue that even if mere machines have a welfare in virtue of having such interests, these interests are also practically irrelevant to our moral deliberations. Therefore, for all intents and purposes (current) machines are not moral patients, or, at least, they are moral patients that we need not care about for their own sake.

I. **Moral Status, Interests, and Welfare**

Before turning to questions concerning the moral status of artificial consciousnesses, it is important to clarify how we are to understand terms such as 'moral status', 'moral patient',

[6] On some views of obligation, we might not have any obligation at all if we can't possibly know what it is; that is all our obligations are evidence and capacity relative. Instead, I will proceed as if our obligations are independent of our epistemic context but that we are excused in circumstances where those obligations are unknowable. See(McMahan 2009), for example, on the distinction between permissibility and excuse.

'interest', and 'welfare.' These terms have intuitive meanings and have a variety of technical meanings in the literature. In what follows, I will define the terms as they will be used below.

There are many sources of *moral status*; entities might matter in our moral deliberations for a variety of reasons, and the domain of things with moral status is extremely large.[7] One kind of moral status, *moral considerability* (Goodpaster 1978; Cahen 2002) or *inherent worth* (Sandler 2007; Sandler and Simons 2012) concerns the moral status an entity has in virtue of having a welfare constituted by interests that are to be taken into account in moral deliberations.[8,9,10] In what follows, I will use the term moral patient to refer to any individual that is morally considerable.[11,12]

Finally, as I will use the terms an individual's *interests* are those things the satisfaction of which contributes to its *welfare* or *well-being*. and the dissatisfaction of which undermine or

[7] Others have used the term 'intrinsic value' to mean what something similar to what I mean by 'moral status' (Floridi 2002). It is worth explicitly saying that this paper is not a defense of the view that machines lack any kind of moral status or intrinsic value. There may be many sources of intrinsic value, however, I argue that they are not moral patients as that term is defined below.

[8] Though these are sometimes conflated (see (O'Neill 2003), inherent worth is here understood to be different than intrinsic value.

[9] We need not take all the interests of all who are morally considerable into account at all times. If a being is morally considerable then we ought to take its interests into account in contexts where we suspect there will be an appreciable impact on that being's welfare.

[10] I remain neutral, as much as possible, on how interests are to be taken into account and weighed against one another. For example, what constitutes equal consideration of interests and what one's various interests entitle one to will differ on deontological or consequentialist views.

[11] This use of 'moral patient' differs from some other uses of the term. The term is, at least sometimes, used synonymously or nearly synonymously with 'moral status' as I have defined it above.

[12] There is an important sense in which being a moral patient is agent relative or, at least, relative to agents of a kind. If there are agents that are radically different than us, for example, they are psychologically incapable of taking the suffering of non-agents into account, then they cannot have obligations to non-human animals in virtue of the interests those non-human animals have in virtue of suffering. For beings like that, it is possible that non-human animals aren't moral patients even while they are for agents like us. The question of how our agency informs which things are patients relative to us is an interesting one, but I set it aside here. My conclusions are intended to apply to agents like us and my claims about which things are patients relative to us.

frustrate its welfare. Below I discuss various kinds of interests an entity may have; having any of those kinds of interest is sufficient for an entity's having a welfare.

There is a longstanding debate about the nature of welfare and the types of interests that constitute such a thing. On some views having a welfare requires that an entity be conscious (Singer 2009; Feinberg 1963; Feldman 2004). If this is so, mere machines, machines that are not conscious, cannot be moral patients. However, on other views, Objective-List Views, consciousness is not always a necessary condition for having a welfare (Streiffer and Basl 2011; Griffin 1988). There is not sufficient space available to adjudicate between these views. In order to argue that current machines are not moral patients on the most charitable understanding of how they might have a welfare, I will assume that an Objective-List View is true and explain how a mere machine might have interests on such a view before arguing that those interests do not justify counting current machines as moral patients.[13]

Below I will distinguish *psychological interests* from *teleo interests*. [14] A psychological interest is an interest that an entity has in virtue of certain psychological capacities (and

[13] On some particular Objective List Views having consciousness will be a necessary condition for having a welfare. On such views, access to the objective goods is only possible for conscious beings. Even on such views, an individual's welfare will not depend soley on his or her particular mental states.

[14] Assuming that an Objective-List view of welfare is true, nothing precludes there being other kinds of interests. It might be, for example, that being green is objectively good for an entity on some view. More plausibly, those that endorse what is often called the "capabilities approach" to well-being argue that certain objective features of a life, like having the resources to pursue projects and the freedom to do so, contribute to welfare independently of any attitudes a given individual may have (see (Sen 1993; Nussbaum 2001)). Another kind of Objective-List view is known as a dignity or integrity view. On such views, it is a component or constituent of welfare that a human or animal's integrity or dignity be maintained or respected. Such views are often appealed to in arguments against the creation of transgenic organisms (Bovenkerk, Brom, and Van Den Bergh 2002; Gavrell Ortiz 2004). In this paper I discuss only psychological and teleo interest. This is because while some components or constituents of welfare, such as those described in the capabilities approach are not strictly psychological, my arguments are intended to show that many of these components of welfare have as a precondition that an entity have certain psychological capacities, namely the capacity for attitudes. With respect to those components or constituents of welfare that do not have psychological capacities as a precondition, such as dignity- or integrity-based accounts of welfare, the arguments against the moral significance of teleo interests of machines apply equally well to these components of

psychological states). A teleo interest is an interest an entity has in virtue of being teleologically organized. In the following section I turn to the question of whether machines with consciousness are moral patients and in virtue of which psychological capacities they are so. Doing so will provide a framework for evaluating the patiency of current machines. In section IV I turn to whether current machines have teleo interests and the implications for their moral patiency.

II. Moral Patiency and Artificial Consciousnesses

a. The Easy Case: Human-Like Consciousnesses

Imagine that an artificial consciousness has been created. This consciousness is very much like ours. It has pleasant and painful experiences, it enjoys or suffers from certain experiences, it has the capacity for imagination, memory, critical thinking, aesthetic and emotional experience, and moral agency. We can even imagine that this consciousness is embodied and goes about the world like we do. On any reasonable normative theory, theory of welfare, and theory of moral considerability, this being will be our moral equal.

This is because whether an individual is a moral patient depends on its capacities; that is the most plausible theories of moral patiency are *capacity-based*. Furthermore, if two beings' are equal in their capacities they are or should be considered symmetrical with respect to moral considerability or their claim to being a moral patient.[15] If tomorrow we cognitively enhanced a chimpanzee so that it is our equal in cognitive capacity, even the most adamant proponent of

welfare. Furthermore, such accounts fail to meet the requirements of non-arbitrariness and non-derivativeness discussed below.

[15] This does not mean that there are no cases where we should favor one life over another. For example, if we must decide between saving the life of our own child and a stranger, we have good reason to save our child. However, this isn't because our child is a more important moral patient. It has to do with the consequences, values, and relationships at stake.

animal experimentation would have to recognize that this chimpanzee deserved protections equal to those afforded other human beings. While being a member of a particular species may matter in moral deliberations, for example, because members of a particular species are endangered or play an important role in an ecosystem, it is difficult to see how being a member of a particular species or other kind is affects whether a being is morally considerable. After all, we have no trouble thinking of alien species that are otherwise like us as moral patients in the same way that we are, and yet they are not of the same species as us. Organisms that are otherwise similar are similarly morally considerable even though we might acknowledge that considerations other than the interests of those beings might favor promoting one being's interests above another's.

Given that the best accounts of moral patiency are capacity-based and the claims about symmetry above, we should recognize that once there are artificial consciousnesses with capacities very much like ours, they will be moral patients, and these patients will be our moral equals. It does not matter that such beings wont' be of our species or that they won't be made of materials similar to ours. In conflicts we will not have priority over such beings.

b. The Hard(er) Case: Animal- and Other Consciousnesses

Questions surrounding the moral patiency of artificial consciousnesses would be very easily answered if we had reason to expect that all such consciousnesses would be very much like us. Unfortunately, depending on how we proceed, there is a much higher probability that in our quest to create artificial consciousness, we will develop consciousnesses that are psychologically more like non-human animals or, that are, psychologically, different than anything that we know of.[16] Therefore, we must venture to discover which capacities in particular give rise to psychological interests.[17]

To determine which conscious machines are moral patients at all, independently of how we are to take them into account, we must first determine which capacities in particular give rise to psychological interests of the sort that are morally relevant. Not all capacities will give rise to morally relevant interests. If we create a consciousness with only the capacity for experiencing colors but with no attending emotional or other cognitive response, we need not worry about wronging said consciousness. It might be a shame to destroy such a consciousness since its creation would no doubt be a fascinating and important achievement, but we would not wrong the machine in virtue of frustrating its interests just as we do not wrong a person (that has consented and is otherwise unaffected) by alternately showing it a red square and then leaving it in darkness.

So, which psychological capacities give rise to psychological interests? To proceed, it is helpful to start by thinking about a non-human animal, say a dog. Hitting such an animal with a sledge hammer is certainly bad for its welfare and at least partly in virtue of the fact that it frustrates its psychological interests. But, in virtue of what are those interests frustrated? Hitting a dog with a sledge hammer causes a variety of psychological phenomena. It causes pain (understood as a sensory experiences), suffering (understood as an aversive attitude towards

[16] Whether the probability of creating consciousness unlike our own is high or low depends on how researchers attempt to create artificial consciousness. If scientists try to simulate human minds by creating functional replicas, then the consciousness created, if such a research program succeeds, is likely to be very much like our own. On the other hand, if scientists try to program or simulate consciousnesses that bear more resemblance to non-human animals or are completely novel, the probability is much higher.

[17] We must also determine how psychological interests of various kinds and strengths should be weighted when they come into conflict. However, since my concern is whether machines are patients at all, I do not address this issue.

some other mental state). It might also result in frustration if it unsuccessfully tries to perform actions that would be possible were it not injured (insofar as dogs are capable of this psychological response). In non-human primates, a strong blow from a hammer might result in all of these plus additional frustration as the primate realizes that its future plans can no longer be realized. Which of these psychological capacities (the capacity for conscious experience of pain, suffering, frustration, future planning) is necessary or sufficient for having psychological interests (that are morally relevant)?

I take it that the mere capacity for sensation is not sufficient to generate psychological interests[18]. We can imagine a being that is capable of feeling the sensations that we call painful but lacking the capacity to have an aversive attitude towards these sensations. If we imagine that sensation is the only cognitive capacity this being has, then this being is very similar to the consciousness that can experience colors. It would not harm this being to cause it to feel those "painful" experiences; such a being just would not care, would not be capable of caring, that it is in such a state. Furthermore, adding capacities such as the ability to recall the sensations won't make those sensations morally relevant.

Of course, we must be careful. It is extremely plausible that our welfare can be improved even if we don't have attitudes one way or another about which state of affairs we are in. Consider two individuals Sam and Sham that have qualitative identical lives in all but one respect; Sam's wife is faithful to him while Sham's wife is secretly adulterous such that Sham will never find out. It seems entirely plausible that Sam has a better life than Sham. This is especially obvious if Sham has a preference that his wife not be adulterous (even though he will

[18] For further argument against this view, called Sensory Hedonism, see (Feldman 2004).

never know that the preference has gone unsatisfied). However, even if Sham were to truly say "I don't care if my wife is adulterous", it seems plausible that Sam has a better life. Authenticity is a welfare-enhancing property of a life (Nozick 1974).[19]

Why not think that a consciousness that can only feel the sensations we associate with pain has an interest in an authentic life? Because, it isn't clear what an authentic life would be for such a being. It seems that the contribution that authenticity makes to welfare, while objective, is a contribution that can only be made to the lives of beings with a certain set of capacities; for example, the capacity to understand authenticity (even if that being doesn't care about it). So, while I'm sympathetic to the idea that there are objective components to welfare, many are only components of welfare for beings with a certain set of cognitive capacities.

While the mere capacity for first-order consciousness or sensory experience is not sufficient for an entity's having psychological interests, the capacity for attitudes towards any such experiences is sufficient. Peter Singer has argued that sentience, understood as the capacity for suffering and enjoyment, is sufficient for moral considerability. Consider the case of newborns and the severely mentally handicapped. These beings are severely limited in their cognitive capacities perhaps lacking all but the capacity for sensory experience and basic attitudes regarding those experiences.[20] And yet, these beings are moral patients. We ought and do take their welfare into consideration in our moral deliberations. We avoid things that

[19] Those that disagree will also be inclined to disagree about the relationship of teleo interests to welfare. Those that reject any non-mentalistic components to welfare will then agree with my assessment that mere machines are not moral patients.

[20] We learn more and more about child consciousness all the time, and so perhaps this is empirically false. However, we don't need to know more about the consciousness of babies to know that they are moral patients.

cause pain in newborns, at least in part, *because* they don't like it or have an adverse reaction to it.

Is having the capacity for attitudes necessary for having psychological interests? That depends on which components of welfare, like authenticity, depend on a being's having psychological capacities. While Sham's life might be improved independent of his particular attitudes about his spouse, could his life be improved by being more authentic if he didn't have the capacity for attitudes at all? Is having the concept of an authentic life sufficient for having a psychological interest in an authentic life? I'm skeptical that it is. One reason for such skepticism is that depending on how we understand what it means to have a concept, it might turn out that computers already possess concepts. Suppose that the artificial intelligence programs used in drones or in self-driving cars make use of concepts.[21] They may have the concept STOPSIGN or TARGET in the sense that they are able to reliably identify or somehow classify things as falling under those concepts. If that's possible, perhaps it is also possible to give machines the concept AUTHENTICITY. However, assuming that such machines lack attitudes, it is rather unnatural to think their existence is made worse off if they fail to have an authentic existence, for example because a machine takes itself to be a self-driving car when it is really a simulation of a self-driving car.[22] At the very least we should be skeptical that on very permissive understandings of what it is to have a concept that having a concept is sufficient for generating a psychological interest.

[21] Thanks to reviewer 1 for these examples.

[22] Perhaps my failure to see how it could be good for such a machine to have an authentic existence just stems my failure to even imagine what it would be like to be a being with no attitudes but with concepts. At the very least, those who wish to disagree owe us an argument that having the concept of authenticity on a very permissive view of concepts is sufficient for having a psychological interests.

Given the above, it seems that any artificial consciousnesses with the capacity for attitudes are moral patients.[23] The same is true for a range of other capacities related to having attitudes such as having desires or preferences. However, assuming that skepticism about the sufficiency of concepts or other mental capacities to generate interests are correct, any machine lacking attitudinal capacities, conscious or otherwise, is a *mere machine*; such machines lack morally relevant, psychological interests. If mere machines have a welfare, it will be in virtue of interests that are not psychological.

c. Epistemic Challenges

Before turning to the question of whether current machines are moral patients, it is important to note some epistemic challenges that we face in determining whether current or future machines have psychological interests. The Problem of Other Minds is the problem of saying how it is we know that other human beings, beings that seem very much like ourselves, have a mental life that is similar to ours.

Perhaps the best answer we can give to this problem is that all our evidence suggests that others are mentally like ourselves. The source of this evidence is evolutionary, physiological, and behavioral. We know that we share a common ancestor with those that seem just like us; we know that they are physiologically like us (and we think we understand some of the bases of our conscious states); and, we know that we behave in very similar ways (for example by avoiding painful stimuli and telling us that such stimuli hurt).

These same sources of evidence can be appealed to in order to justify claims about the mental lives of animals. The more closely related (evolutionarily) and the more physiologically

[23] I'm assuming that any individual that has the capacity for attitudes has at least one attitude about something.

and behaviorally similar an organism is to us, the better our evidence that it has cognitive capacities like ours[24]

Unfortunately, in the case of machines, we lack the typical sources of evidence about their mental life. A computer lacks any evolutionary relationships with other species, its physiology is entirely different than any other conscious being's, and if it feels pain, cannot tell us it feels pain or behave in a way that suggests that it is in pain unless it has been somehow enabled to do so. Unless we have a very good understanding of the (functional) bases of mental capacities and so can know whether such bases exist in a given machine, we may be largely in the dark as to whether a machine is conscious and as to whether it has the morally relevant capacities described above.

I do not have any solutions to the epistemic problems raised by the nature of machine consciousness. However, these difficulties do raise ethical concerns. As we get closer to creating artificial consciousness it will be important to examine these difficulties very carefully to make sure we can distinguish mere machines from those machines with psychological interests. To fail to do so, might put us in a situation where we create and potentially torture beings that deserve our moral respect when this could be avoided.

The question remains, do these concerns apply to current machines and if so, which ones? If they do, our current attitudes and behaviors towards machines may be entirely inappropriate.

[24] There is considerable controversy over which mental capacities non-human animals have. See (Tomasello and Call 1997) for a discussion of some of the issues concerning primate cognition. However, there is little doubt that many non-human animals have aversive attitudes towards those conditions we identify as painful. See (Varner 1998, chap. 2) for an overview of the evidence that non-human animals have the capacity for suffering.

III. Mere Machines

Fortunately, the ethical concerns just discussed apply to current machines only if those machines are not mere machines. In the remainder of this paper, I hope to argue that current machines are mere machines and that, even though they may have a welfare in virtue of having non-psychological interests, they are, for all practical purposes, not moral patients. In this section, I take up the claim that current machines are mere machines.

Consider our most advanced computers, from chess computers, to missile guidance systems, to IBM's Watson. We have absolutely no evidence that such computers are conscious and so absolutely no evidence that such computers have the capacity for attitudes that would ground psychological interests. Of course, we could program a computer to tell us that it doesn't like certain things, I'm sure even Apple's Siri has "attitudes" of this sort. But, we know that such behaviors are programmed and we don't believe that computers genuinely have cognitive capacities on these grounds.

One could of course argue that we have no evidence the other way either. An expert on the neurological bases of cognitive capacities might try to respond by saying that the functional bases for the relevant capacities, as best we understand, are not realized in any machines or computers that currently exist. I am no such an expert and so will offer no such argument. Instead, let me grant that we have no evidence either way. Of course, it would also seem to follow that we don't have evidence either way whether toasters or corkscrews have these capacities.

If the correct attitude to have regarding current machines is agnosticism, doesn't this provide reasons to be skeptical of my claim that current machines are mere machines?

Technically, yes. However, everyone should agree that, even if today's machines have psychological interests, we have little or no idea what will promote or frustrate those interests, which experiences they enjoy or are averse to, and no way to discover them. After all, since we have no evidence about the cognitive capacities of machines whatsoever, we have no reason to believe that, for example, a computer enjoys doing as it was programmed to do as opposed to hating doing those things.

Given where agnosticism leads, for all intents and purposes today's machines are, I argue, mere machines. To see why, we must first recognize the distinction between permissibility and excuse (McMahan 2009). Whether an act is permissible depends only on whether we have an obligation not to perform it; if we have no such obligation not to perform some act, then that act is permissible. Obligation and, thereby, permissibility are a function of the moral facts in a given context independently of the epistemic situation of the agent who will perform the action. Excusability, on the other hand, is a function of the epistemic situation of an agent. An agent can be excused, that is, not held responsible, for an act that is impermissible.

To further illustrate the distinction between permissibility and excuse, assume for the moment that utilitarianism is true and some agent, A, is deciding between buying a car and not buying a car. The agent considers all the information available to her to the best of her ability. She performs a utilitarian calculation on the basis of these considerations and decides to purchase the car. After purchasing the car, she takes a road trip to Washington D.C. and accidentally hits and kills the President of the United States who, unexpectedly, ran into the middle of the street. This causes the stock market to plummet and, for dramatic effect, let's say it starts a series of wars. We can also stipulate that had the agent not purchased the car none of this would have occurred. From a utilitarian perspective, purchasing the car was impermissible since it did not maximize utility. However, given that the agent considered the outcomes of her decision to the best of her ability and made an honest mistake, she should hardly be held accountable; she is excused for acting impermissibly.

Given the distinction between permissibility and excuse and the fact that even if current machines do have psychological interests we do not have any sense of what they are, we can see why we can behave as if current machines are mere machines. It may be that there is, in an objective sense, a moral obligation to take into account the psychological interests of current machines on the assumption that they have them, but our inability to do so in any reasonable way excuses us from any obligations we violate. If we can't possibly determine our obligations, we are surely excused for failing to live up to them.[25] Until the day that we can determine whether current machines are conscious or have good reason to think we may be creating artificial consciousnesses and some idea of what their attitudes are, we ought to behave as if current machines are mere machines.[26]

IV. Teleo Interests

Given the argument of the previous section, unless we have reason to believe that current machines have a welfare in virtue of having non-psychological interests, current machines are

[25] If we understand obligations as being a function of our epistemic context so that what's obligatory and permissible is limited by what we can or can't know, the argument is even easier to make. By *ought implies can*, we can only be obligated to take the psychological interests of machines into account if it were possible to determine what those interests were. But, given our current limitations we can't make such a determination. Therefore, we are under no obligation whatsoever to take current machines into account and may permissibly behave as if they are mere machines.

[26] The alternative is to give up research involving machines. Until we have good reason to believe that we are creating the functional bases for consciousness, considering a ban on machine research seems overly restrictive.

not moral patients. In Section I, I explained that on some views of welfare, Objective-List Views, consciousness is not a necessary condition for having a welfare. On such views, an entity can have interests that are totally divorced from and independent of mental life. In the remainder of the paper, I consider whether mere machines have what I take to be the most plausible kind of non-psychological interests.

a. The Interests of Non-Sentient Organisms

It seems obvious that there are ways to benefit or harm non-sentient organisms. Pouring acid on a maple tree is bad for it, providing it with ample sunlight and water is good for it. So, there is intuitive plausibility to the idea that such beings have interests. However, proponents of views on which consciousness is a necessary condition for having a welfare have long denied that such beings have a welfare and that statements about what is good for or bad for non-sentient organisms is either incoherent (Singer 2009) or reduce to claims about what is good for sentient organisms (Feinberg 1963).[27,28] For example, such philosophers might argue that the reason that acid is bad for maple trees is that we have a preference for the growth of maple trees flourishing, and so it is bad for us if we were to pour acid on them.

[27] Important to the debates about the coherence of attributing interests to non-sentient organisms is the distinction between taking an interest and having an interest (See (Taylor 1989; Varner 1998; Basl and Sandler Forthcoming; Basl and Sandler Forthcoming) on the distinction). While taking an interest in X certainly would seem to require consciousness, since it implies caring or otherwise having some attitude about X, proponents of the interests of non-sentient organisms argue that something can have an interest in X, X can be good for that thing, independently of the interests it has. A common, though controversial, example might be an interest that smokers might have in giving up smoking independently of whether they actually care about doing so.

[28] It's worth noting that proponents of views on which what's good for non-sentient organisms are derivative on our interests can't easily account for at least some ascriptions of interests to non-sentient organisms. For example, weed killer is instrumentally valuable for us precisely because it is bad for weeds. It would be strange to say that weed killer is good for weeds; it is good for killing weeds. However, this worry is not decisive; the way we talk is, at best, a starting point for thinking about these issues.

In order to respond to such arguments, proponents of the welfare of non-sentient organisms must explain how these being's interests aren't merely a product of our anthropomorphizing, how our attributions of interest to these beings aren't mere metaphor; they must explain how the interests of non-sentient organisms are non-derivative and non-arbitrary. If there is no basis for the interests of such organisms except in virtue of the interests of sentient organisms or our arbitrarily deciding that what's good for us is good for them, then we should regard non-sentient organisms as lacking interests.

The most prominent and promising attempt to meet the challenges of derivativeness and arbitrariness is to ground the interests of non-sentient organisms in their being goal-directed or teleologically organized.[29] Non-sentient organisms have parts and processes that are organized towards achieving certain ends such as survival and reproduction. There is a very real sense in which, independent of our desires, maple trees have parts and processes whose goal or end it is to aid in survival and reproduction in the ways characteristic of maple trees. A maple tree is defective if it fails to grow leaves, in part because it is the end of certain sub-systems to produce leaves and promote growth.

Given that organisms are teleologically organized, it is possible to specify a set of interests, non-arbitrarily and non-derivatively, in light of this organization. Whatever promotes the ends of organisms is in its interest, whatever frustrates those ends undermines its interests.[30] These are often referred to as *biological interests* (Varner 1998), but I will call them

[29] All existing organisms will have interests of this kind, but sentient organisms will have additional interests.
[30] A similar account of the interests of non-sentient organisms can be found in (Varner 1998).

teleo interests in virtue of the fact that they are grounded in the teleological organization of organisms and not, strictly speaking, their being biological.

Some might balk at the notion that plants are genuinely teleologically organized. In a world that was not designed, where organisms have evolved through a combination of chance processes and natural selection, how it is possible that organisms without minds could have ends? The answer is that natural selection grounds claims about such ends. It is perfectly legitimate to ask what makes for a properly functioning heart as opposed to a defective one. The answer to such a question is that a defective heart fails to pump blood. But, this can only be a defect if the heart has an end or purpose. And, it does; the purpose or end of the heart is to pump blood, as opposed to making rhythmic noises, because that is what it was selected for.[31] Natural selection serves as the basis for teleology. It does so not because it is a directed process with aims or ends, it is not, but because natural selection explains, in terms of selection of an organism's ancestors, why it is that organisms have the traits that they do and thereby what it is that the traits are there for – i.e. they are there to perform or do that which they were selected for.[32,33]

b. Derivative Interests

While various environmental ethicists have been keen to adopt the view that natural selection grounds the teleological organization of non-sentient entities and thereby the sole interests of such beings, they have been adamant that artifacts do not have a welfare (Goodpaster 1978;

[31] This is a very brief summary of what is known as the etiologically account of functions (Wright 1973; Millikan 1989; Neander 1991; Millikan 1999; Neander 2008).

[32] Of course, not all traits are the result of selection. They may be the result of drift, or an evolutionary spandrel (Gould and Lewontin 1979). Those that wish to ground teleology in natural selection need not be adaptationists.

[33] Whether past selection explains why a given organism has the trait that it does is a matter of some controversy. See for example (Sober 1984; Neander 1988; Forber 2005).

Varner 1998; Taylor 1989). This is strange because while some may balk at thinking that organisms are teleologically organized, there is no denying that machines and most other artifacts are teleologically organized. Even if natural selection cannot ground teleology, the intentional creation of an entity can. The parts of my computer have purposes and the computer itself myriad ends. Why then shouldn't we judge that artifacts have interests?

There are various differences between organisms and machines. The former are biological, more natural, etc. However, none of these are morally relevant differences between the two, and none of these differences explain why teleological organization gives rise to interests in organisms but not artifacts.[34] One difference that many have thought is morally relevant has to do with the nature of that teleological organization; the teleological organization of artifacts, it is often said, is derivative on our interests, while the interests of organisms is not derivative.[35] Therefore, the only sense in which mere machines and artifacts have interests is derivative. Call this objection the Objection from Derivativeness.[36]

The Objection from Derivativeness is mistaken. First, let's carefully distinguish two reasons we might think that the so-called welfare of mere machines is derivative. The first reason is that mere machines only exist for our use. If we had no need, desire or use for them, mere machines would not exist. Call this *Use Derivativeness*. The second reason is that the ends or teleological organization of mere machines can only be explained by reference to the

[34] For a discussion of these issues see (Basl and Sandler Forthcoming; Basl and Sandler Forthcoming)

[35] Another way to conceptualize this difference is as a difference in the nature of the selection processes that give rise to or explain organisms and artifacts. Artifacts are the result of artificial selection processes while organisms are the result of natural processes.

[36] For a more detailed discussion of this objection and others, as well as a more rigorous defense of the application of the etiological account of teleology to artifacts see (Basl and Sandler Forthcoming; Basl and Sandler Forthcoming).

intentions or ends of conscious beings; the explanation of the teleological organization in mere machines is derivative on our intentions. Call this *Explanatory Derivativeness*.

Being Use-Derivative is not an obstacle to being genuinely teleologically organized or to having interests. Many non-sentient organisms, from crops to pets, are use-derivative and yet they still have teleo-interests. It would be bad for a field of corn to suffer a drought even if that field had been abandoned. If we decided to clear a forest and replant it, the plants that grew would have the same interests as the plants that grew there before (assuming they are the same species). The fact that mere machines exist to serve our purposes makes it such that what promotes their ends is typically the same as what promotes our ends, but this fact doesn't undermine the idea that there are things that promote the *machine's* ends. It is still the subject of teleo interests even if it wouldn't have those interests if not for us.

The same is true concerning explanatory derivativeness. The fact that we must appeal to another's intentions to explain the teleological organization of a machine does not show that the machine is not teleologically organized, that it does not have its own ends. Were I to have a child and play an influential role in his or her life and career choice, it would matter none at all to whether a promotion benefitted the child. Even though, perhaps, you could not explain the preferences my child will have without reference to my intentions, the child is still has interests of its own. The same is true of interests grounded in teleological organization. Despite the fact that you must cite a designer's intentions to explain why a machine has the ends that it does, it, the machine, still has those ends.

Furthermore, proponents of the legitimacy of teleo interests in non-sentient organisms cannot appeal to explanatory-derivativeness to distinguish non-sentient organisms from mere machines. The evolutionary history of many non-sentient organisms involves the intentions of a variety of sentient beings. We have every reason to believe that the explanations for why many organisms are organized as they are would be incomplete if they did not refer to the intentions of sentient beings. It is because early hominids had certain intentions that modern dogs are as they are, and it is likely that many plant species evolved in response to the intentions of non-human primates. So, many non-sentient organisms have welfares, in virtue of having biological interests, that are explanatorily-derivative. This does not undermine their having a genuine welfare. Or, if it does, it also undermines biocentrism entirely.

c. The Interests of Mere Machines

The account of teleo interests described above is the most plausible way of grounding claims about the welfare of non-sentient organisms. The Objection from Derivativeness constitutes the best objection to the claim that it is only non-sentient organisms and not machines or artifacts that have teleo interests. Given the failures of the objection, the following thesis should be accepted:

> **Comparable Welfare Thesis**: If non-sentient organisms have teleo interests, mere machines have teleo interests.

This principle does not commit us to the view that either non-sentient organism or mere machines have a welfare, nor does it commit us to the view that non-sentient organisms have a welfare if mere machines do. However, for those sympathetic to the idea that non-sentient organism have interests and those interests constitute a welfare, the principle commits us to expanding the domain of entities that we recognize as having a welfare to include machines and artifacts.

Some will not be convinced by the arguments above and will deny that machines have teleo interests in the way that organisms do. Others will maintain that neither non-sentient organisms nor non-sentient machines have teleo interests. I will not here provide any further argument that non-sentient organisms have teleo interests, nor will I provide any independent arguments that mere machines have teleo interests. My goal is to show that even given that machines have a welfare grounded in teleo interests, we need not concern ourselves with their welfare. In what follows, I will assume that mere machines have teleo interests and ask what this means for our moral deliberations.

V. The Practical Irrelevance of Teleo Interests

Finally we turn to the question of the moral relevance of teleo interests. There are some arguments to the effect that teleo interests are not interests in the sense that their satisfaction contributes to welfare (Behrends 2011; Rosati 2009). According to these arguments, teleo interests pick out *relative goods*, things that are good for an organism only relative to some end. On such an understanding, to say that some resource R is good for some subject S is just to say that R achieves some end for S, but nothing more. This provides grounds to distinguish relative goods from those goods or interests relevant to welfare. If this is right, and since being a moral patient requires having a welfare, mere machines are not moral patients.

Even if relative goods are welfare goods, there is good reason to think that mere machines are not moral patients, or, more precisely, that for all practical purposes they are not moral patients; we need not worry about their teleo interests in our moral deliberations. This is for at least two reasons. First, most often, since we wish to use artifacts as they are intended, our use is in accordance with their interests. Using a machine to serve the purpose for which it was intended does not result in a conflict of interests.[37]

Second, even in circumstances where our actions would frustrate a machine's teleo interests, our legitimate psychological interests always take precedence over teleo interests.[38] To see that this is so, consider cases where a conflict arises between the teleo interests of an individual that also has psychological interests, a human being. A human's teleo interests will include the proper functioning of their organs and other biological systems, but often humans have preferences that these systems fail to work properly. For example, an individual that does not desire to have children might take steps to impair their biological capacity to do so. In such a case, there is a conflict between teleo interests, the interests associated with proper functioning of reproductive parts, and psychological interests, the attitudes and preferences regarding offspring. In this case, psychological interests take precedence and it is morally permissible to frustrate the teleo interests in this case.

Some people attribute significant importance to reproductive capacities and so might not be convinced by this case. Besides, one might argue that the number of biological interests that would be frustrated is smaller in number than the psychological interests that would be frustrated by disallowing this biological process. In order to establish the priority of

[37] This may be true only if we also do what is necessary to maintain a machines capacity to serve that purpose. It is possible to be hard on machines. For example, we can brake too hard and too often in our car. In doing so, while we use the braking system for the end for which it was designed, we also undermine the brakes capacity to continue to serve that purpose. Thanks to Jeff Behrends for pushing me on this point.

[38] It is not that teleo interests are never relevant or that we never desire to promote them. When someone is in a coma, for example, the best we can do, often, is to help satisfy their teleo interests. However, this is not a case of a conflict between psychological and teleo interests.

psychological interests, a case is needed where it is morally permissible to satisfy a legitimate psychological interest at the cost of even a very large number of teleo interests.

Consider a case involving non-sentient organisms. Imagine that biologists were convinced that there were something of importance to be learned by first growing and then ultimately destroying a very large number of trees (say 1 million). Let's imagine that it would teach us something about the origins of plant life on earth. Assuming no negative externalities, this experiment seems permissible. This is so despite the fact that a massive number of teleo interests would be frustrated, and the only immediate gain would be the satisfaction of a psychological interest we have in learning about the origins of our world.

This shows that legitimate psychological interests trump teleo interests even when the number of teleo interests frustrated is very large. But, in almost all cases where there will be a conflict between our psychological interests and a machine's interests, our psychological interests will be legitimate; we will be frustrating machine interests to gain knowledge about machines and to develop new ones and to improve our well being. For this reason, there seems to be no problem now, or in the future, with our frustrating the teleo interests of mere machines either by destroying them or otherwise causing them to function improperly. You may recycle your computers with impunity.

Before concluding it is worth briefly discussing three objections. The first is that the above argument doesn't show that mere machines are never moral patients for practical purposes, only when the psychological interests they conflict with are *legitimate*.[39] There are cases where the teleo interests of mere machines might make a difference to our moral

[39] Thanks to Ron Sandler for pushing me on this point.

deliberations, where those interests cannot be entirely discounted. These cases might involve the wanton destruction of non-sentient organisms or machines.

These cases should not worry us very much. Firstly, there are very few who wish to, for example, destroy a million computers or trees for no good reason. Secondly, such acts, in practical circumstances, would be morally wrong for many reasons since, for example, they would have many negative externalities. In cases where the moral wrongness of an act is overdetermined it is hard to tell whether any of the wrongness results from the patiency of the individuals whose interests are frustrated by the act. Furthermore, since the wrongness of such acts is overdetermined, we need not worry, for practical purposes about the patiency of those with teleo interests.

A second objection might be that teleo interests might always be trumped by psychological interests, but there are, potentially, conflicts that involve only teleo interest.[40] For example, the conflict between a parasitic plant and its host, for example, is a conflict only between teleo interests. Certainly these kinds of conflicts can arise in the case of machines, as when a chess computer plays against itself.

While these kinds of conflicts may arise, they are not conflicts that are morally interesting because they involve no agent. Consider a conflict between a zebra and a lion. This is a genuine conflict of interests, but if agents can do nothing about the conflict it is not a conflict for which it makes sense to ask “what ought we to do?” So, the conflicts of interests that are of moral concern, that is the conflicts of interest that agents should concern themselves with adjudicating, are those conflicts that involve agents. But, as soon as an agent

[40] Thanks to an anonymous reviewer for raising this issue.

considers how it is that they should adjudicate a conflict, the question will arise as to which psychological interests of the agent or others are at stake in the conflict, thus introducing an interest that I've argued, will trump any biological interests at stake.

I suppose we can imagine a case where an agent is observing two machines engaged in some conflict and where there are no legitimate psychological interests at stake. For example, an agent who cares very little about the outcome of a competition between two chess computers might ask "given the teleo interests of these machines, should I intervene on behalf of one machine or another?" But, this is not a situation we find ourselves in. We have chess computers that play against one another for legitimate scientific and educational ends.

A third objection might deny my claim about the thought experiment involving the million trees. Someone might argue that our interest in evolutionary knowledge does not justify the destruction of 1 million trees even if there are no additional externalities. They might accuse me of begging the question in my defense of the prioritization of psychological interests over teleo interests.

There is little that can be said here. There is no thought experiment that will not beg the question. I take it that research involving machines, even the destruction of machines, is unproblematic even when the psychological interests at stake aren't very strong. In light of this, we need not be sensitive to the interests of machines. But, there is little more I can say here that would convince those that fundamentally disagree about the value of such research.

Conclusion

The arguments of the previous section temper any worries we may have about the moral wrongs we might commit against mere machines. In the near future, no matter how complex the machines we develop, so long as they are not conscious, we may, as far as concerns the artifact itself, do with them largely as we please.[41] However, things change once we develop, or think we are close to developing artificial consciousness.

Once artificial consciousnesses exist that have the capacity for attitudes they have psychological interests that ground their status as moral patients. We must, at that point, be careful to take their welfare into account and determine the appropriate way to do so. And, given the epistemic uncertainties surrounding the creation of consciousnesses and the nature of their psychological interests, we must proceed with care as we create machines that have what we think are the functional bases of consciousness.

Works Cited

Basl, John, and Ronald Sandler. Forthcoming. “The Good of Non-Sentient Entities: Organisms, Artifacts, and Synthetic Biology.” *Studies in History and Philosophy of the Biological and Biomedical Sciences*

———. Forthcoming. “Three Puzzles Regarding the Moral Status of Synthetic Organisms.” In *“Artificial Life”: Synthetic Biology and the Bounds of Nature*, ed. G. Kaebnick. Cambridge, MA: MIT Press.

Behrends, Jeff. 2011. “A New Argument for the Multiplicity of the Good-for Relation.” *Journal of Value Inquiry* 45 (2): 121–133.

Bovenkerk, B., F. W. A. Brom, and B. J. Van Den Bergh. 2002. “Brave New Birds: The Use of’Animal Integrity’in Animal Ethics.” *The Hastings Center Report* 32 (1): 16–24.

Cahen, Harley. 2002. “Against the Moral Considerability of Ecosystems.” In *Environmental Ethics: An Anthology*, ed. Andrew Light and H. Rolston III. Blackwell.

Feinberg, Joel. 1963. “The Rights of Animals and Future Generations.” *Columbia Law Review* 63: 673.

Feldman, F. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties and Plausibility of Hedonism*. Oxford University Press, USA.

Floridi, Luciano. 2002. “On the Intrinsic Value of Information Objects and the Infosphere.” *Ethics and Information Technology* 4 (4): 287–304.

Forber, Patrick. 2005. “On the Explanatory Roles of Natural Selection.” *Biology and Philosophy* 20 (2-3): 329–342.

[41] Of course, we have many reasons to take artifacts into account for other reasons. They may be other people’s property, they may be valuable scientific achievements, etc.

Gavrell Ortiz, Sara Elizabeth. 2004. “Beyond Welfare: Animal Integrity, Animal Dignity, and Genetic Engineering.” *Ethics & the Environment* 9 (1): 94–120.

Goodpaster, Kenneth. 1978. “On Being Morally Considerable.” *The Journal of Philosophy* 75: 308–325.

Gould, Stephen J., and Richard Lewontin. 1979. “The Spandrels of San Marcos and the Panglossian Paradigm.” *Optimizing Learning and Evolutionary Change in Behavior* 153.

Griffin, James. 1988. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford University Press, USA.

McMahan, Jeff. 2009. *Killing in War*. 1st ed. OUP Oxford.

Millikan, Ruth Garrett. 1989. “In Defense of Proper Functions.” *Philosophy of Science* 56 (2): 288–302.

———. 1999. “Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function.” *The Journal of Philosophy* 96 (4): 191–206.

Neander, Karen. 1988. “What Does Natural Selection Explain? Correction to Sober.” *Philosophy of Science*: 422–426.

———. 1991. “Functions as Selected Effects: The Conceptual Analyst’s Defense.” *Philosophy of Science* 58 (2): 168–184.

———. 2008. “The Teleological Notion of ‘Function’.” *Australasian Journal of Philosophy* 69 (4) (March 24): 454 – 468.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Nussbaum, Martha C. 2001. *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

O’Neill, John. 2003. “The Varieties of Intrinsic Value.” In *Environmental Ethics: An Anthology*, ed. Holmes III Rolston and Andrew Light.

Rosati, C.S. 2009. “Relational Good and the Multiplicity Problem1.” *Philosophical Issues* 19 (1): 205–234.

Sandler, Ronald. 2007. *Character and Environment: A Virtue-Oriented Approach to Environmental Ethics*. Columbia University Press.

Sandler, Ronald, and Luke Simons. 2012. “The Value of Artefactual Organisms.” *Environmental Values* 21 (1): 43–61.

Sen, Amartya. 1993. “Capability and Well-Being.” In *The Quality of Life*, ed. Amartya Sen and Martha Nussbaum, 1:30–54. Oxford University Press.

Singer, Peter. 2009. *Animal Liberation: The Definitive Classic of the Animal Movement*. Reissue. Harper Perennial Modern Classics.

Sober, Elliott. 1984. *The Nature of Selection*. MIT Press Cambridge, MA.

Streiffer, Robert, and John Basl. 2011. “Applications of Biotechnology to Animals in Agriculture.” In *The Oxford Handbook of Animal Ethics*, ed. T Beauchamp and R Frey. Oxford.

Taylor, Paul W. 1989. *Respect for Nature*. Studies in Moral, Political, and Legal Philosophy. Princeton, N.J.: Princeton University Press.

Tomasello, Michael, and Josep Call. 1997. *Primate Cognition*. 1st ed. Oxford University Press, USA.

Varner, Gary. 1998. *In Nature’s Interest*. Oxford: Oxford University Press.

Wright, Larry. 1973. “Functions.” *Philosophical Review* 82: 139–168.