# WILL INTELLIGENT MACHINES BECOME MORAL PATIENTS?

Parisa Moosavi

York University

**Abstract:** This paper addresses a question about the moral status of Artificial Intelligence (AI): will AIs ever become moral patients? I argue that, while it is in principle possible for an intelligent machine to be a moral patient, there is no good reason to believe this will in fact happen. I start from the plausible assumption that traditional artifacts do not meet a minimal necessary condition of moral patiency: having a good of one's own. I then argue that intelligent machines are no different from traditional artifacts in this respect. To make this argument, I examine the feature of AIs that enables them to improve their intelligence, i.e., machine learning. I argue that there is no reason to believe that future advances in machine learning will take AIs closer to having a good of their own. I thus argue that concerns about the moral status of future AIs are unwarranted. Nothing about the nature of intelligent machines makes them a better candidate for acquiring moral patiency than the traditional artifacts whose moral status does not concern us.

## 1. Introduction

Recent advances in Artificial Intelligence (AI) and machine learning have raised many ethical questions. A popular one concerns the moral status of artificially intelligent machines (AIs). AIs are increasingly capable of emulating intelligent human behaviour. From speech recognition and natural-language processing to moral reasoning, they are continually improving at performing tasks we once thought only humans can do. Their powerful self-learning capabilities give them an important sense of autonomy and independence: they can act in ways that are not directly determined by us. Their problem-solving abilities sometimes even surpasses ours. What's more, they are taking on social roles such as caregiving and companionship, and thereby seem to merit a social and emotional response on our part. All this has led many philosophers and technologists to seriously consider the possibility that we will someday have to grant moral protections to AIs.[1] In other words, we would have to "expand the moral circle" and include AIs among *moral patients*, i.e., entities that are owed moral consideration.[2]

This question of moral patiency is the focus of my paper. Roughly speaking, the question is whether future AIs will be the kind of entities that can be morally wronged and need moral protection. My position, in a nutshell, is that concerns about the moral status of AI are unjustified. Contrary to the claims of many authors (e.g., Coeckelbergh 2010; 2014; and Gunkel 2018; 2019; Danaher 2020), I argue that we have no reason to believe we will have to provide moral protections to future AIs. I consider this to be a commonsense view regarding the moral status of AI, albeit one that has not been successfully defended in the philosophical literature. This is the kind of defence that I plan to provide in this paper.

## 2. The Status of Moral Patiency

We may start by clarifying the concept of moral patiency further. Frances Kamm's account of this moral status is particularly helpful. According to Kamm (2007, pp. 227-229), an entity has moral status in the relevant sense if *it counts morally in its own right and for its own sake.* Let's consider each of these conditions in turn: (i) what it is for an entity to *count morally*, (ii) what it is for it to count morally *in its own right*, and (iii) what it is for it to count morally *for its own sake*.

To say that an entity *counts morally* is to say that it is in some way morally significant. More specifically, it means that there are ways of behaving toward the entity that would be morally problematic or impermissible. An entity that counts morally gives us moral reasons to do certain things and act in certain ways toward it, such as to treat it well and not harm it. We typically consider humans as entities that count morally in this sense, and ordinary rocks as entities that do not. But almost anything can count morally in the right context. If an ordinary rock, for instance, is used as a murder weapon and becomes a piece of evidence, it may be morally impermissible to temper with it.

There are, however, different ways to count morally, not all of which amount to being a moral patient. An entity can count morally, but merely instrumentally so—i.e., because our treatment of it has a morally significant effect on others.. To have the relevant moral status, the

[1] See, e.g., Coeckelbergh (2014); Schwitzgebel and Garza (2015); Gunkel 2018); and Gordon (2020).

[2] See Singer (1981) on the idea of expanding the moral circle.

entity in question must count morally *in its own right*, i.e., non-instrumentally. In other words, it must be valued *as an end*, and not merely as a means. The above-mentioned rock does not meet this condition, but our fellow humans do: no further end needs to be served by the way we treat them for us to have a reason to treat them well.

Moreover, an entity with moral patiency counts morally *for its own sake*, which is to say that we have reason to treat it in a certain way for the sake of the entity itself. Note that an entity might be valued as an end but not for the sake of itself. For instance, the aesthetic value of *Mona Lisa* can give us reason to preserve it independently of the pleasure or enlightenment it can bring. This, however, does not mean that we have reason to preserve *Mona Lisa* for the sake of the painting itself. We do not think of preservation as something that is good *for the painting*. We rather think we have reason to preserve it because the painting has value *for us*. We value the painting as an end, but—to borrow Korsgaard's term—this non-instrumental value is still "tethered" to us: we are the beneficiaries of this value.[3] In contrast, our moral reasons to save a human being from drowning are reasons to do something for *their* sake. They get something out of being saved that *The Mona Lisa* does not.

Thus, on Kamm's account, an entity has moral patiency when it can give us reason to treat it well, independently of any further ends that such a treatment might serve, and precisely because being treated well is good for the entity itself.

This is the conception of moral patiency that I will adopt going forward.[4] The question I am interested in is, therefore, whether future AIs will be the kinds of entities that can give us moral reasons of this specific kind: reasons that have to do with *what is good for the entity itself*. I am not concerned with whether there will be other sorts of moral reasons to treat them in a certain way. I am only asking whether they will qualify for moral patiency proper.

The next section offers a critical review of the existing literature on the moral status of AI, and argues that it does not provide a satisfactory answer to the question at hand.

## 3. The Question of Moral Patiency Applied to AI

[3] See Korsgaard (2018, pp. 9-15).

[4] This conception, or a sufficiently similar one, is broadly shared among philosophers discussing moral patiency (e.g., Singer 1973; 1981; 1993; Varner 1998; Taylor 1986; Basl 2014; 2019; Basl and Sandler 2013).

There are two challenges that make our question a difficult one. The first has to do with the general question of what grounds moral patiency, and the second specifically concerns AI.

Regarding the general question, the various accounts on offer disagree on what criteria are primarily relevant to moral status and why. On broadly Kantian accounts, for instance, moral status depends on rational capacities like autonomy and practical reasoning.[5] In contrast, broadly utilitarian accounts consider the capacity to desire or to experience pleasure and pain to be most relevant.[6] Other, less predominant accounts consider entering certain types of social or biological relationship to be necessary or even sufficient for moral patiency.[7] It is especially hard to adjudicate between these accounts because there is also a lot of pre-theoretical disagreement about the particular cases we might use to test them. There is, for instance, a lot of disagreement about the moral status of fetuses, humans in a permanent vegetative state, many non-human animals, and the environment.

The case of AI also presents a special challenge of its own, which has to do with the fact that AIs are constantly changing and acquiring new capabilities. Note that it is not really *current* AI systems that we are most worried about. The real question is whether *future* AIs will qualify for moral patiency, and we simply do not know what the future holds for AI. Experts do not agree on what level of intelligence AI might be able to ultimately achieve.[8] We do not know what characteristics future AIs will have, especially with respect to the capacities that figure in the predominant accounts of moral patiency. Note that capacities like autonomy, rationality, consciousness, and sentience are themselves notoriously difficult to pin down. And although there is a good deal of work on whether current cases of concern such as non-human animals and human fetuses at various stages of development have the relevant capacities, we are not in a position to

[5] On Kant's view, rational capacities are both the source of moral obligation for moral agents and the feature of moral patients that grounds their dignity and the respect they are owed. See Kant (1785, pp. 434, 436) on the relation between autonomy, dignity, and respect. For contemporary versions of the view, see Korsgaard (1996a; 1996b), Wood (1998), O'Neill (1998), and Regan (2002).

[6] These capacities are often considered prerequisites for having interests that need to be incorporated into the utilitarian calculation. See Singer (1973; 1993) for a prominent example.

[7] These accounts differ widely regarding what kind of relationship they consider relevant and whether they consider it to be necessary, sufficient, or both. See Nozick (1997), Callicott (1989), Anderson (2004), and Warren (1997) for different examples.

[8] See Grace et al. (2018) and Müller and Bostrom (2014) for recent surveys of expert opinion of the future of AI.

recognize these capacities in a radically different form. So, if future AIs were to acquire the relevant capacities, it is not clear how we would be able to tell.[9]

The literature on the moral status of AI takes two broad approaches to addressing this latter challenge, which we may call the *speculative* and the *revisionist* approaches. The first approach avoids having to make predictions about future AIs by focusing on a hypothetical question about some imaginary form of AI. Rather than asking whether AIs *will* qualify for moral patiency, the authors taking this approach ask whether AIs *would* qualify *if* they were to possess certain capacities. These authors tend to rely on or defend the predominant views about the grounds of moral status. The criteria they consider to be relevant to moral patiency include the capacity for consciousness (Andreotta 2020; Mosakas 2021; Johnson & Verdicchio 2018) and phenomenal desire (Novelli 2020) to autonomy (Tonkens 2012; Gordon 2021) and moral agency (Sullins 2006; Gordon 2020). While some of these authors presuppose specific accounts of moral status, others rely on fewer assumptions about what grounds moral patiency. Schwitzgebel and Grazia (2015), for instance, leave it open which exact psychological and social properties are relevant to moral status, but argue that AIs *would* qualify for moral patiency *if* they achieved a 'human-level' degree of the relevant psychological and social properties.

The speculative approach allows these authors to sidestep questions about the indeterminate future of AI while still making interesting claims about the relevance or irrelevance of certain aspects of AI to moral patiency. However, to the extent that their arguments avoid questionable assumptions, they do little to inform our present and future decisions about *actual* AIs, which have no demonstrated connection to the imaginary forms of AI they hypothesize. As noted, the properties that figure in standard accounts of moral patiency are themselves difficult to pin down or predict. So, knowing that AIs would be moral patients *if* they had some of these properties does not exactly tell us what we should do. Some authors make the seemingly stronger claim that it is *possible* for AI to acquire the relevant properties, in that it would not be inconceivable or

[9] For example, there is no agreed-upon philosophical theory of consciousness, and the various tests offered for machine consciousness have serious limitations. From Alan Turing's famous Turing Test (Turing 1950) to Susan Schneider's more recent and sophisticated AI Consciousness Test (Schneider 2019), many such tests focus on behavioural indistinguishability. While this allows them to be relatively neutral among competing theories of consciousness, it also raises serious questions about whether passing the test is sufficient for having consciousness (see Udell and Schwitzgebel 2021).

inconsistent with the nature of the relevant properties for AIs to acquire them.[10] However, this mere logical or metaphysical possibility does not offer any more practical guidance in orienting ourselves toward future AIs.

The second approach gets around the difficulty of assessing the moral status of future AIs by revising the criteria that are standardly taken to be relevant to determining moral patiency. The authors taking this approach tend to reject what they call the ‘standard’ or ‘properties-based’ view, according to which moral status depends on whether an entity possesses certain intrinsic or ontological properties. They argue that we can determine how we should treat AIs entirely based on extrinsic or relational criteria, which are philosophically and empirically easier to pin down. Coeckelbergh (2010; 2014) and Gunkel (2018; 2019), for instance, propose “a relational turn” and a “paradigm shift in moral thinking”. On their view, rather than asking what ontological features AIs will possess, we can determine their moral status by asking what relations they will have with us and each other. The moral status of an AI can thus be decided based on empirically observable criteria, like how people bond with it and whether they treat it as a fellow companion or a mere machine. Another example of the revisionist approach is Danaher’s (2020) defence of “ethical behaviorism”. Danaher argues we can determine how we should treat AIs entirely based on behavioural evidence of their performance. On his view, we should treat AIs as moral patients if and when they are “roughly performatively equivalent” to other entities we consider moral patients. Danaher’s view is different from those defending a relational account of moral status in that he remains agnostic on whether behavioural indistinguishability is sufficient for *having* moral patiency. He rather argues that evidence from behaviour is sufficient for treating AIs *as if* they are moral patients, because it is the only epistemic ground available to us. That said, he similarly rejects the idea that intrinsic or ontological properties are relevant to our treatment of AIs.

If successful, the revisionist approach would be much more practically informative regarding our question. The proposed criteria are much less epistemically challenging to determine and predict. In fact, some of the authors taking this approach suggest that current AIs are well on their way to meeting the criteria.[11] Nevertheless, the shift away from the standard accounts of moral patiency is a radical and contentious step that needs to be independently motivated. And as many

[10] See, e.g., Schwitzgebel & Grazia (2015, pp. 103-108).

[11] Danaher (2020), for instance, argues that today’s AIs may be already part of the way there in achieving behavioral “performative equivalence” with other moral patients.

critics have argued, the arguments in favour of this shift tend to fall short. The advocates of the "relational turn" like Coeckelbergh (2010; 2014) and Gunkel (2018; 2019) argue against the property-based accounts by pointing out that these accounts face epistemological challenges and are anthropocentric in that they start from properties that humans exemplify. However, it is far from clear that these concerns justify the conclusion that the moral status of an entity does *not* depend on its ontological features. As Mosakas (2020) and Müller (2021) have argued, solely relying on how *we* relate to an  entity leads to a subjectivist form of relativism that is at least *prima facie* suspect, not to mention more problematically anthropocentric. Danaher's (2020) argument for "ethical behaviorism" similarly cites epistemic limits as the reason for watering down the grounds for treating an entity as a moral patient. He argues that we normally make inferences about moral status on the basis of behaviour alone, because there are no other grounds for attributing moral status that are both relevant and epistemically accessible. However, the fact that *we* normally attribute moral patiency based on behaviour alone is hardly sufficient to show that such inferences are always valid. And as I argue later in §8, the assumption that we do not have epistemic access to any other relevant facts is simply false.

Thus, both approaches face problems in answering our question. The speculative approach avoids making controversial assumptions about the grounds of moral patiency at the cost of offering little guidance on how to orient ourselves toward future AIs. And the revisionist approach provides practical guidance only by weakening the criteria of moral patiency in questionable ways. My aim below is to offer an answer to the question that is based on minimally controversial assumptions about moral patiency and yet affords meaningful guidance regarding future AIs.

## 4. A Minimal Necessary Condition for Moral Patiency

As we saw, even the standard, property-based accounts of moral status disagree about what criteria are primarily relevant to moral patiency. Rather than adopting any of these accounts and assuming that any specific capacity like consciousness, sentience, autonomy, or rationality is necessary for moral patiency, my strategy will be to start with a minimal necessary condition that directly follows from the concept of moral patiency presented earlier. Since my aim is not to offer an account of the grounds of moral patiency but merely to assess the case of AI, I will keep my substantive commitments to a minimum.

Following Kamm's account, we saw that a moral patient *counts morally in its own right and for its own sake*: it gives us reason to treat it in certain ways independently of any further ends that such a treatment might serve for other entities, and precisely because being treated in that way is good for that entity itself. What follows is that for an entity to be a moral patient, it must be such that things can be good or bad *for* it; it must be the kind of entity that can benefit or be harmed by our treatment. Thus, simply considering the concept of moral patiency provides us with a necessary condition: to be a moral patient, the entity in question must have *a good of its own*, in reference to which states of affairs can be said to be good or bad for it. Rocks or paintings do not meet this condition, as they do not have a stake in how things turn out, whereas human beings do, since states of the world can be better or worse for them.

This necessary condition on moral patiency is minimal in that it does not make assumptions about what the good of the entity consists in and whether the entity must have any cognitive, affective, or agential capacities. Note that different substantive accounts of what is non-instrumentally good for a beneficiary entity (which is also referred to as *well-being* or *welfare*) come with different commitments regarding what capacities the entity must have. Hedonist theories, which claim that well-being consists in the greatest balance of pleasure over pain, imply that the entity must have the capacity to experience pleasure or pain to have a good at all.[12] In contrast, Objective List theories, which claim that what is non-instrumentally good for a person includes items from a list of objective goods, do not necessarily imply that.[13] That is because on these views the goodness of the items on the list does not depend on whether the entity derives pleasure or satisfaction from them. That said, even on Objective List theories, the beneficiary entity must somehow be the *receiving subject* of the goods on the list. To say, for instance, that knowledge and achievement are non-instrumentally and objectively good does not mean that just any entity (e.g., a rock) benefits from their existence in the world. The goodness of the goods must be somehow tied to the beneficiary entity. Thus, even if the goodness of the goods turns out to be *objective* and independent from the subjective attitudes of the beneficiary entity, it must be still *subject-relative*, i.e., good *for* the beneficiary entity.[14] And depending on the nature of the goods on the list, this subject-relativity might require certain capacities, e.g., the capacity for knowledge, on the part of the beneficiary subject.

[12] For contemporary versions of hedonism, see Feldman (2004) and Crisp (2006).

[13] See, e.g., Finnis (1980); Griffin (1986); and Fletcher (2013).

[14] See Sumner (1996, pp. 21-22) on the subject-relativity of well-being.

Some authors maintain that to be a beneficiary entity at all, an entity must be a *subject* in a more robust sense that requires the capacity for subjective experience. Sumner (1996), for instance, argues that Objective List theories fail to account for the subject-relativity of well-being and that subject-relativity requires a beneficiary entity to have “a reasonably unified and continuous mental life” (p. 43). Similarly, Korsgaard (2013) argues that what explains the *relational* character of goodness—i.e., that what is good is good *for* an entity—is the fact that there are conscious organisms that can perceive things and respond positively or negatively. On the opposing side, there are authors who maintain that the capacity for subjective experience or consciousness is not necessary for having a good of one’s own. Most notably, proponents of biocentrism in environmental ethics argue that even nonsentient living organisms can have a good of their own and thus be candidates for moral patiency (see, e.g., Goodpaster 1978; Attfield 1981; Taylor 1986). On their view, all living organisms—whether sentient or nonsentient—are goal-directed, teleological systems with natural purposes of their own; and these purposes underwrite their ‘biological interests’ and make them recipients of benefits and harms. My aim here is not to settle the question of whether having a good of one’s own requires consciousness, sentience, or a mental life—the way I construe the minimal necessary condition on moral patiency, it is silent on this question. However, since I plan to argue *against* ascribing moral patiency to AIs, I will take the biocentrists’ claim seriously and consider the possibility that some entities might be capable of having a good of their own without having consciousness or sentience. In other words, for the sake of argument, I will grant that at least in the case of AI subjects, having a good of one’s own might be possible without consciousness or sentience.

Having identified a minimal necessary condition for moral patiency, we can now ask whether future AIs will meet this condition. Before addressing this question, however, it would be helpful to examine the case of nonintelligent artifacts and machines, which are the very first predecessors of AI. The next two sections look at these simpler artifacts and machines, which I will group together and refer to as *traditional artifacts*.

## 5. Do Traditional Artifacts Have a Good?

Most of us consider it obvious that simple artifacts like knives and chairs, or even more complex machines like cars and radios, do not have a good of their own. Although we sometimes speak in terms of what is ‘good’ or ‘bad’ for these artifacts and machines, we do not take such claims literally. We might say, for instance, that changing a car’s oil frequently is good for the car, but we do not mean that the car genuinely benefits from having its oil changed. What we mean is that changing the oil is good for the *functioning* of the car, i.e., serving its function successfully for a longer time.

As obvious as this may sound, we need to examine its rationale, especially if we are to take seriously the possibility that having a good does not require the capacity for consciousness or sentience. We saw that many biocentrists claim even nonsentient organisms can be ascribed a good of their own because they are goal-directed, teleologically organized entities. This teleology-based argument is particularly relevant to the case of artifacts, because artifacts too are goal-directed, teleologically organized entities. A knife, e.g., is for the purpose of cutting, and is structured with parts and aspects that are each designed to perform a specific function enabling it to serve this purpose. If a nonsentient organism like a tree can be ascribed a good of its own in virtue of its functional organization, one might wonder why the same thing cannot be said about knives or other traditional artifacts.

The teleology-based argument for biocentrism takes various forms (see, e.g., Goodpaster 1978; Taylor 1986; and Varner 1998). But the general idea is that, whether sentient or nonsentient, living things have natural purposes of their own. They are directed toward maintaining and reproducing themselves, and are functionally organized in a way that enables them to reach these goals. The roots of a tree, for instance, have the function of absorbing and transporting water and nutrients from the soil, and are directed toward enabling the tree to maintain itself. These natural purposes and functions give us a nonarbitrary basis for identifying what is good or bad for an organism in terms of what is conducive or detrimental to the functioning of the organism's parts. We can say, for instance, that watering a tree is good for it, and cutting its roots off harms it.

Some critics of biocentrism have argued that if being teleologically organized was a basis for attributing a good of one's own, the scope of entities with a good of their own would include traditional artifacts, which also have a teleological organization (Regan 1976; Sapontzis 1987, Basl 2014; 2019; Basl and Sandler 2013). This potential implication is often considered fatal to the teleology-based argument because we do not normally think that artifacts have a good of their own or can genuinely receive benefit or harm from our treatment.[15] Biocentrists have thus tried to block this implication by identifying a relevant difference between organisms and artifacts. Most

[15] Notable exceptions are Basl (2019; 2014) and Basl and Sandler (2013), who consider the implication that artifacts have interests of their own counterintuitive but ultimately palatable.

notably, they have argued that the teleology of living things is *their own*, whereas as the teleology of artifacts and machines is somehow borrowed, or derivative of *our* goals, intentions, and interests. Only original, nonderivative goals can underwrite the ascription of a good of one's own, so only organisms can be ascribed such a good on that basis (Arbor 1986; Goodpaster 1978; Taylor 1986).

This response, however, is criticized for not clarifying what makes the goals of artifacts and machines derivative and why that matters. As Basl and Sandler (2013) have argued, some organisms are designed by synthetic biologists in a similar manner to artifacts. The teleology of these *synthetic* organisms seems to be just as derivative as that of artifacts, yet this does not mean that they do not have a good of their own. Basl and Sandler consider two different senses of derivativeness, and argue that the teleology of synthetic organisms is derivative on either interpretation. If derivativeness means an entity only exists because it serves a purpose for us, they argue, this is also true of synthetic organisms like crops and farm animals that we have created for our own use. If, on the other hand, derivativeness means the functional organization of an entity can only be explained by reference to our intentions and goals, then again there are synthetic organisms whose traits and characteristics are shaped by us. Basl and Sandler thus argue that regardless of how or why the functional organization of an entity is shaped the way it is, it is the entity's *own*, just as a human agent's goals are her own regardless of how they are acquired or whether others have had a role in shaping them.[16]

Basl and Sandler are right that the difference in the causal histories of organisms and artifacts does not necessarily amount to a relevant difference in their teleology. An entity's being causally shaped by intentional, rather than natural, selection does not by itself disqualify it from having goals of its own. To see this, note that we normally don't think whether living things are intentionally created by God makes a difference to whether they have genuine goals or a good of their own. However, I will argue that there is another sense of derivativeness that characterizes the teleology of artifacts and is relevant to the question at hand. What I have in mind has less to do with artifacts' *causal* dependence on us and more to do with their *constitutive* dependence. In the next section, I will give an account of this constitutive sense of derivativeness and explain why it is relevant to the question of having a good.

[16] See also Basl 2019, pp. 130-160.

## 6. The Derivate Nature of Artifact Teleology

To see the relevant notion of dependence, note first that teleological claims are inherently normative. A goal or function is not just something that a goal-directed entity or function-bearing part of a system does, or does regularly, but something that it is in some sense *supposed to* do. The entity can *succeed* or *fail* in attaining its goal, and the parts can perform their functions *well* or *poorly*.

Note, further, that artifacts typically have various causal capacities besides the capacity or capacities that constitutes their function. A butter knife, e.g., can reflect light and collect dust in addition to its capacity to spread butter. But unlike the latter, reflecting light or collecting dust are not the knife's function: we do not evaluate the knife based on how well it performs with respect to these capacities. It seems straightforward that part of what singles out spreading butter as the knife's function is the fact that *we* find this capacity useful, and *we* conceptualize the knife as an artifact that is supposed to serve this purpose. In other words, it is our goals, intentions, and beliefs that—at least in part—determine the knife's function. Without us, even if the knife were to somehow pop into existence and have the same causal capacities, it would not have that *function*.

To be clear, I am not here talking about the role of our intentions and beliefs in making the physical entity that is the artifact. The point is not that without our contribution an artifact would not physically exist or have its causal capacities—though that may very well be the case. It is rather that without a suitable relation to our intentional states, an artifact's causal capacities would not *count* as functional. In fact, our constitutive role in determining an artifact's functions can arguably be established without a causal role. Although most paradigmatic instances of artifacts are made by us, an artifact does not have to be physically made by humans to have a function. As McLaughlin (2001) has argued, for instance, a log lying across a creek can have the function of being a bridge even if it actually fell there in a storm, e.g., if we *decided* to leave it there to use it for crossing the creek (p. 60). Neither is having a causal role in shaping something sufficient for conferring a function on it. If we were to accidentally knock down a tree during construction, causing it to fall over the same creek, our causal role alone would not be enough to confer a function on it. Whether something that we have created has a function depends on the mental event that accompanies its creation.[17] How our intentional states come to be embodied in an artifact, and

[17] See McLaughlin 2001, pp. 45-46.

which intentions and beliefs determine an artifact's function, varies from case to case. The artifact may be designed and manufactured with a particular purpose in mind, or it may be later adapted for a different use, earning it a different function. But regardless of the details, the artifact needs to have the right relation to *someone*'s intentional states to qualify as having a function at all. As McLaughlin (2001) puts it, "the truth conditions for artifact function ascriptions involve the beliefs and desires of agents" (p. 60).

In my view, this constitutive dependence on our intentional states captures the relevant sense in which artifact functions are derivative. As Thompson (2008) makes the same point, the teleology of artifacts is characterized by a kind of "partial idealism": the truth of teleological judgments about artifacts "presupposes that someone makes or has made the corresponding judgment, or at least some others belonging to the same system of judgments" (p. 80). It simply cannot be true that "butter knives have a blade *in order to* spread butter" unless someone has made this judgment or at least some other judgments regarding what knives are supposed to do. In contrast, at least intuitively, the teleology of organisms does not seem to be derivative in this sense. The ends that we ascribe to organisms and the functions we ascribe to their parts and aspects do not seem to depend on how we view them or whether we take an interest in them. As Thompson points out, unrecognized forms of life are common, but an unrecognized artifact is not really an artifact but "a merely possible one" (2008, p. 80).

We can now see that the case of synthetic organisms does not undermine the proposed difference between organisms and artifacts. Our intentional states are *causally* implicated in how synthetic organisms are created and designed. But whether their teleology is owed to us in the constitutive sense is another matter. Of course, to the extent that these organisms are appropriated to serve certain ends for us, they do have derivative functions that are imposed on them by us. But this merely reflects the fact that synthetic organisms are simultaneously organisms and artifacts. It does not mean that they have no teleology of their own in addition to the derivative teleology that we impose on them. Consider, e.g., a domestic breed of sheep that is specifically designed by us to have fine wool best suited to creating textile. The sheep's wool can be ascribed the derivative function of generating textile due to its relation to our ends. But many of the sheep's other parts—e.g., its heart or eyes—can still have non-derivative functions whose status as function is independent of us. In fact, even the sheep's wool may additionally have a non-derivative function like keeping the sheep warm. On the view being developed here, whether our intentions are causally involved in creating an entity is simply *irrelevant* to having non-derivative teleology. The claim is not that an entity's being the product of intentional design *prevents* it from having non-derivative teleology, but that intentional design alone is not sufficient for conferring non-derivative teleology. So, while the fact that we have intentionally created synthetic organisms does not confer non-derivative teleology on them, it does not rob them of non-derivative teleology that they may otherwise have. And to the extent that we can assume synthetic organisms are still *organisms*, it can be argued that at least some of their parts and aspects will have functions independently of how we view them and what interests us about them.

Note that I am not claiming *all* products of synthetic biology are necessarily organisms or have nonderivative teleology. As Basl (2019) has argued, some synthetic organisms are made entirely from scratch, and synthetic biology can in principle build entities with entirely novel genomic sequences (p. 136). We have no reason to assume all such radically novel entities must be organisms. What constitutes a living organism is a difficult question, and in the case of many products of synthetic biology there will be a real question whether they are in fact organisms or mere artifacts. But these hard or marginal cases need not concern us here. My argument does not rely on the claim that it is always easy or even possible to recognize living organisms as such. Basl and Sandler's argument against the relevance of derivativeness appeals to the case of synthetic *organisms*: it relies on the assumption that the synthetic entities in question are in fact organisms and have a good of their own. I have argued that our causal role in creating these organisms does not automatically prevent them from having non-derivative teleology, because derivativeness is not a matter of causal dependence.

The notion of derivativeness as constitutive dependence explains why the derivative teleology of artifacts is not a basis for ascribing a good to them. The ends that we ascribe to artifacts are not really their own, and owe their teleological status to the ends of another entity external to them. So, to the extent that the attainment of these ends amounts to a good, this good also belongs to the external source. In contrast, organisms have ends that are not derived from any external end but are rather internal to them. And to the extent that the attainment of these ends amounts to a good, this good must belong to the organism itself. As Nicholson (2013) has argued, this difference between internal and external teleology also explains a crucial dissimilarity between organisms and artifact when it comes to function ascriptions. Artifacts are ascribed functions, but organisms are not (unless they are also artifacts). It is only parts of organisms that are ascribed functions, whereas in artifacts, both parts and wholes are ascribed functions. According to Nicholson, this is because ascribing a function to an entity implies that the beneficiary of its operation is another entity.[18] An artifact is ascribed functions, because its operation is viewed as benefitting an external agent. The parts or traits of an organism are also ascribed functions, because their operation is viewed as benefitting the organism of which they are a part. But the operation of an organism is not viewed as benefitting anything outside the organism, so it is not ascribed a function.

It may be objected that although the proposed difference tracks our intuitions about artifacts and organisms, the basis for drawing the distinction is not clear. On what grounds do we ascribe *non*-derivative teleology to organisms, and why can't we ascribe the same kind of teleology to artifacts? In fact, it seems that consulting the best philosophical accounts of biological function would undermine the proposed distinction. According to the widely-respected *etiological* account of biological function (Wright 1973; Millikan 1989; and Neander 1991), the function of a trait or a part of an organism is, roughly, the contribution for which it was preserved under the past operation of natural selection. The etiological function of a dog's heart, e.g., is to pump blood, because pumping blood has contributed to the natural selection of dogs with hearts in the past, and thus explains the fact that dogs currently have hearts. Now, one could argue that artifacts have a history of selection just as organisms do, and their selection history similarly explains the presence of their features. A butter knife's capacity to spread butter, e.g., explains why its features have been selected and put together the way they are. If a selection history can ground non-derivative teleology in the case of organisms, it seems that it can do so in the case of artifacts as well.[19]

The problem with this suggestion, however, is that etiological functions are also derivative in the constitutive sense, not only in the case of artifacts but also in the case of organisms. This may sound surprising, because etiological function ascriptions appear to be grounded in objective facts about natural selection. After all, whether a given effect of a function-bearer has contributed to its selection is an objective matter. However, note that function ascriptions are more than just a descriptive claim about what has helped the function-bearer get here. They also have a normative dimension: they imply that the performance of the function-bearer can be evaluated in terms of its function. But just because a trait has an effect that explains how it got here, it doesn't follow that

[18] See also McLaughlin (2001, pp. 140-161).

[19] Thanks to John Basl for drawing my attention to this objection.

its performance should be evaluated based on whether it continues to have that effect. It is only if we attach some value or significance to the effect that explains why the trait got here that we would view it as a *function* and its absence as a *malfunction*. To see this, note that the conditions for etiological function ascription can be met by other structures that we do not describe in functional terms. As Bedau (1991, pp. 651-654) has argued, there are nonliving populations of simple replicating molecules such as crystals forming from clay, which meet the conditions of evolution by natural selection. Although these molecules make contributions that explains their presence, we do not ascribe a function to them, nor do we *evaluate* their performance based on whether they continue to do the same thing.[20]

So, why do biologists ascribe etiological functions to the past contribution of a biological trait to natural selection? What is the value or significance that underwrites the normative character of etiological function ascriptions? The answer, I believe, lies in *our* interest in explaining how organisms have evolved over time. It is in the context of this (cognitive) interest that defining the 'success' or 'failure' of a trait based on its contribution to natural selection makes sense: it provides a convenient way to highlight the effects of a trait that explains precisely what evolutionary biologists are interested in explaining. It is not a coincidence that etiological function ascriptions are primarily used in evolutionary biology, which is focused on offering historical explanations of how populations of organisms evolve. A trait's past contribution to natural selection happens to be the relevant explanatory factor if that is our explanatory project. As many philosophers of biology have argued, in other areas of biology where historical explanation is not the main concern, function ascriptions are not captured by the etiological account, and are primarily tied to the current effects of a trait rather than its evolutionary past.[21] Thus, although facts about natural selection are very much objective and independent of our intentional states, their relevance to ascribing functions is owed to us and our explanatory interests. This is why etiological functions are constitutively derivative and cannot be a basis for ascribing nonderivative teleology or a good of one's own, either in the case of organisms or in the case of artifacts.

[20] Bedau uses this example to argue against the etiological account by showing that the conditions specified in this account are not sufficient for ascribing functions. My aim here, however, is not to criticize the etiological account, but to argue that the concept of function that it introduces is constitutively derivative.

[21] See Walsh and Ariew (2010); Roe and Murphy (2011); and Kraemer (2014).

Basl (2019) has offered an etiological account of welfare that is worth noting in this relation. According to this account, the good of an organism consists in the promotion of its ends, where to be an end is to have been the product of natural selection at the level of individual organism (p. 81).[22] Basl thus maintains that etiological function ascriptions do in fact support the inference to the ascription of a good, not only for organisms but also for artifacts. On his view, the selection history that underwrites these function ascriptions also underwrites the ascription of ends, which in turn underwrites the ascription of a good. However, by defining ends in terms of selection history, Basl is introducing a technical notion of an 'end', and consequently a technical notion of a 'good', which do not correspond to the morally significant notions that concern us here. Note that the talk of *selection* in natural selection is metaphorical, as there is typically no agential choice involved in the process. A trait's being naturally 'selected' merely means that it has proliferated via the differential survival and reproduction of organisms with different traits. A history of natural selection is, thus, not a basis for ascribing ends in the intuitive, morally significant sense. Basl argues that defining ends in this way is "nonarbitrary", because it grounds ends in an objectively definable process (i.e., natural selection) rather than an *ad hoc* attribution of an end or a good (p. 76). However, although the process of natural selection is nonarbitrary with respect to the explanatory aim of evolutionary biology, it is completely arbitrary outside this explanatory context. There is no reason to assume that what historically explains how an entity got here should be considered an *end* in a sense that we should care about in our moral deliberation. In fact, Basl himself acknowledges that his etiological account does not capture a morally significant welfare, either in the case of artifacts or in the case of nonsentient organisms (see pp. 161–182).[23]

Note, however, that this doesn't necessarily mean that organisms do not have nonderivative teleology, only that they do not have it simply in virtue of their history of natural selection. The view that organisms are naturally directed toward ends such as survival and reproduction is independently plausible, and it is not my aim here to either defend or reject it. But whatever the basis for ascribing nonderivative ends to organisms, it cannot merely consist in the fact that they

[22] Basl defines ends directly in terms of the entity's selection history rather than making reference to etiological functions in order to bypass debates on whether the etiological account offers the correct account of the concept of *function*. He also limits the ascription of ends to entities that have been selected at the level of individual organisms to avoid questions about whether the organism as a whole can be said to have the ends of its parts and aspects (see Basl, 2019: 77-82).

[23] See also McShane (2021) for a critical discussion of an etiological account of welfare, especially pp. 3505-3507.

have an evolutionary history.[24] Similarly, we can see that the selection history of artifacts is not sufficient grounds for ascribing nonderivative teleology to them or concluding that they have a good of their own.

Thus, I have argued that there is no basis for ascribing a good to artifacts, even if we do so in the case of nonsentient organisms. Of course, I have not argued that nonsentient organisms do in fact have a good of their own. Nor have I argued that artifacts *cannot possibly* have a good. But the idea that simple everyday artifacts and machines do not have a good of their own is independently plausible, and I have argued against what I take to be the most compelling reason to question this assumption, i.e., the fact that they have a teleological organization. In the next section, I turn to the case of *intelligent* machines, and ask whether we should think that the addition of intelligence will take these artifacts closer to having a good of their own. To do this, I look at the distinctive process by which machines equipped with AI change and improve themselves: *machine learning*.

## 7. Adding Intelligence to Traditional Artifacts

The notion of intelligence is broad and what exactly it takes for a machine to acquire *real* intelligence is debatable. But AI researchers distinguish between "strong" AI, which aims at developing machines that have real or human-level intelligence, and "weak" AI, which aims at developing machines that can perform tasks that require intelligence when done by humans. Strong AI seeks to create artificial persons with the full range of human mental capacities, including phenomenal consciousness. And it is *at best* a long-term prospect. Weak AI, on the other hand, is the more familiar kind of AI that focuses on performing problem-solving or reasoning tasks using algorithms and methods that emulate or augment human intelligence. It should be clear that it is not strong AI but weak AI that concerns us here. As explained in §3, we are not trying to answer the hypothetical question whether AIs *would* become moral patients *if* they acquired the same morally relevant capacities as humans. We are trying to see if we should think AIs *will* become moral patients in the future. To answer this question, we need to look at what we know about AI and what it has achieved so far, and ask whether that should make us think that further progress in its methods and algorithms will result in machines that have a good of their own.

[24] I will say more about our epistemic basis for ascribing nonderivative ends to organisms in §7.

Undoubtedly, AIs are advancing fast, and already have an impressive ability to perform tasks we once thought only humans could do. Moreover, they have an increasingly high degree of independence and autonomy, and can behave in novel ways that go beyond our direct design and ability to predict and control. These capabilities are largely due to machine learning, which is the area of AI that is concerned with creating algorithms that can increase their knowledge and improve their performance over time. Every computer program receives some data as input, processes the input using an algorithm, and produces some result as output. But unlike traditional programs, machine-learning programs are not given a fully specified and fixed algorithm for generating results based on input data. Instead, they have a mechanism that enables them to learn from examples of successful performance. They can, for instance, receive the input data and the desired result, and produce the algorithm that would turn one into the other. In this sense, they are algorithms that "make other algorithms" or "write their own programs" (Domingos 2015, p. 6). Because of this, machine-learning programs can solve many problems that traditional programs have not been able to solve, including ones that even humans struggle with. They can learn to do complicated tasks like classifying images or driving cars without needing step-by-step instructions from us, and they can develop novel methods of accomplishing these tasks.

It is not difficult to see why it can seem that with further advances in machine learning AIs could come to have a good of their own and ultimately attain moral patiency. Unlike traditional artifacts, whose goals and functions depend on how we value or perceive them, AIs seem to break away from our influence and even our understanding. Their ability to learn from experience and modify their own algorithms seems to enable them to acquire new goals and functions that are not determined or predicted by us. And it can appear that someday they will acquire enough independence that we simply will have to view their goals and functions as *their own.* We can see this line of thought in Danaher's (2020) remarks about machine learning and its potential to take AIs beyond serving the ends of their human creators. He notes, for instance, that "certain robotic manufacturing processes—particularly those that incorporate machine learning—may result in robots that do not serve any clearly interpretable end or an end that is readily associated with their original creators." He thus argues that we should view AIs not merely as entities designed for the purpose of serving us, but "much more like humans who have been loosely programmed by evolution and cultural development". Basl similarly points out that recent developments in machine learning can create "dynamic, self-maintaining systems, capable of unpredicted, novel behavior" via processes very similar to natural selection (Basl 2019, p. 154).[25] Basl uses an example from Bostrom (2016) to illustrate how an AI's goals can move in a direction that is not anticipated or desired by us. He invites us to consider an imaginary machine-learning algorithm that is extremely skilled and tasked with creating paperclips. While gradually improving and optimizing its paperclip-making algorithm, the paperclip-maker might learn that it needs to safeguard its power supply and thus generate defence mechanisms to make sure that it never shuts down. Or it might predict that the best source of some key material is derived from a resource that humans heavily rely on, and as a side effect it may destroy us by consuming resources essential to our survival. It seems evident that at that point, the paperclip-maker is pursuing its own goals independently of what we value or are even aware of.

However, this perception of the effect of machine learning is mistaken. A closer look reveals that although machine learning changes the teleology of AIs in a way that reduces their causal dependence on us, it does not make any difference to their constitutive dependence. And the new goals that can result from machine-learning processes are not the machine's own any more than those of a traditional artifact are the artifact's own. As we saw in §5, a goal is not just something an entity does regularly, but something it is *supposed to* do—something that sets the standard for evaluating its success or failure. That the paperclip-maker starts to safeguard its own power supply or consumes certain resources does not imply that these new tendencies have the status of goals. Insofar as they do have this status, it is in virtue of their instrumental relation to what we already consider a goal of the machine: making paperclips. It is only to the extent that safeguarding the power supply is a *means* to the AI's ultimate *end* of making paperclips that it qualifies as a goal. And the basis for attributing *that* end to the AI is not any different than it would be in the case of a traditional, non-intelligent paperclip-making machine: its relation to *our* goals, intentions, and beliefs. Thus, even when machine learning changes the teleology of an AI, the newly acquired goals are merely *intermediate* goals that ultimately owe their normative character to us. They are, therefore, just as derivative as the original goals the machine started with. In fact, when an AI like Bostrom's paperclip-maker starts to exhibit behaviours that conflict with our interests or pose risks for us, it is not obvious that the new behaviours should even qualify as intermediate goals. A more fully specified description of the AI's intended ultimate goal—e.g., making paperclips *safely* or

[25] Basl does not think that AIs or other artifacts qualify for moral patiency. But he believes that artifacts do have a good of their own, and uses the example of AI make this idea more plausible.

*without using too much power*—would clarify that not just any behaviour that helps maximize the number of paperclips would count as an intermediate goal for the AI. In some cases it might be more apt to characterize an AI's learned behaviour in terms of a malfunction or failure rather than a newly acquired goal.

It is worth noting that current AIs already have the ability to identify effective means to their assigned goals and pursue these means as new intermediate goals. In fact, they can do this without our knowledge and in such a way that it is hard for us to interpret or identify their intermediate steps. Consider, for example, Deep Learning, which is currently the most prominent and widely successful method in AI. Deep Learning algorithms have a complex architecture consisting of a network of nodes and connections with very many tweakable parameters. Each of these elements serve internal functions within the overall network, and together they create the mechanism that generates an output based on the algorithm's input. During its training phase, the network is given huge high-dimensional datasets, which are not necessarily in a format that makes sense to us. In a network tasked with labelling images, for instance, the training data often consists of stored images that are represented by matrices of numerical RGB values for the pixels. As the network trains itself, its parameters change, and its elements acquire new internal functions. For example, in trying to detect images of cats, a certain group of nodes might start to detect whether a picture contains two sharp edges. These intermediate steps, however, will not necessarily be known to us, as often even experts cannot make sense of what the parameters of the system mean and why they yield a particular output from a given input.[26] Thus, AIs equipped with Deep Learning already possess the feature of Bostrom's futuristic paperclip-maker that was supposed to imply it has a good of its own. If acquiring novel or unexpected intermediate goals was a basis for attributing a good, we would have to count many instances of currently existing AIs as already having passed the threshold. However, as I argued above, these newly acquired goals are derivative: their status as goals ultimately depends on us. So, regardless of the machine's degree of independence from our knowledge or causal influence in acquiring them, they do not make a difference to whether it has a good of its own. What intelligence and learning capabilities afford AI is better means-end reasoning: the ability to recognize and take the necessary steps to performing a task effectively. And we can expect that future AIs will hone their capacity for instrumental reasoning further. They

[26] This is why Deep Learning algorithms are said to be "opaque" or like a "black box". See, e.g., Burrell (2016) and Weller (2017).

will master the art of planning ahead, predicting the utility gained from each step, and developing new strategies for solving a problem. But there is no reason to think that improving at instrumental reasoning would take them closer to having nonderivative goals or a good of their own.[27]

Of course, this is not to say it is *impossible* or *inconceivable* for AIs to form nonderivative or, say, start genuinely caring about a particular outcome. My claim is not that something about the nature of intelligent machines categorically precludes them from ever acquiring a good of their own or moral patiency.[28] I have rather argued for the more modest claim that we have no reason to believe we should attribute a good of their own to them now or in the future. To the extent that our only reason for attributing goals and functions to AIs is their connection to our intentions, goals, and perceptions, they are not relevantly different from traditional artifacts. And there is no reason to believe further progress in the methods and algorithms of AI will change that.

One might object that there *is* a reason to believe it will: we can expect the behaviour of intelligent machines to become increasingly complex and humanlike, to the point that there will be no relevant basis for distinguishing them from organisms that we already view as having nonderivative goals, a good of their own, or even moral patiency. The manifest behaviour of intelligent machines is, after all, very different from that of traditional artifacts. Simple inanimate artifacts like knives and chairs do not exhibit any behaviours we associate with goal-directedness or other morally relevant capacities like sentience or autonomy. Intelligent machines on the other hand already exhibit complex and spontaneous behaviours, and we can expect that in the future their behaviour will resemble that of humans and nonhuman animals even more. If one day their behaviour becomes indistinguishable from that of the paradigmatic cases of moral patiency, how can we justify treating them differently? It seems we will have no epistemic basis for denying that they too have whatever capacities are necessary for moral patiency, be it having a good of their own, or other capacities like sentience or autonomy.[29]

[27] See Bostrom (2012) for a related discussion on the irrelevance of intelligence to nonderivative or “final” goals.

[28] As I mentioned earlier, for instance, my argument is different from Bryson’s (2010) claim that AIs cannot have moral patiency simply because they are owned and designed by humans.

[29] In fact, even the degree of intelligence exhibited by today’s AIs has been enough for some to suspect that AIs have already acquired some of these capacities or are well on their way to doing so. The concerns raised by Google engineer Blake Lemoine about the possible sentience of a natural-language-generation program, LaMDA, is a recent example (see Lemoine 2022a; Lemoine 2022b; Tiku 2022).

Note that this objection need not claim that moral patiency or the capacities that are relevant to having it simply consist in exhibiting a pattern of behaviour. What the objection targets is our *epistemic* basis for attributing the relevant capacities to the paradigmatic cases of moral patients but not behaviourally indistinguishable AIs. The objection thus presents a version of the kind of methodological behaviourism that Danaher (2020) defends. In §3, I classified Danaher's "ethical behaviourism" alongside revisionist views aiming to replace the standard criteria of moral patiency with more epistemically accessible criteria. But it should be noted that unlike other revisionists like Gunkel and Coeckelbergh, Danaher does not deny that possessing certain ontological properties may very well be necessary for moral patiency. He does not claim that moral patiency can be reduced to exhibiting a certain pattern of behaviour, but merely that given our epistemic limitations, observable behaviour is the only available ground for attributing patiency.

The objection from ethical behaviourism essentially bypasses my argument in this section. I have argued that examining the mechanism that turns traditional artifacts into intelligent machines—i.e., machine learning—does not give us reason to believe future AIs will become moral patients. The objection, however, contends that if instead of focusing on the mechanism of change we look at the end result, we do find such a reason. The next section addresses this objection.

## 8. Why Intelligent Behaviour Is Not Reason Enough

According to Danaher, we should treat AIs as moral patients if and when they become "roughly performatively equivalent" to other entities we consider to be moral patients (2020, p. 2024). Danaher acknowledges that many people will find this standard counterintuitive and think there are other epistemically accessible facts about AIs that tell against ascribing moral patiency to them. So, he considers a number of these potential "epistemic defeaters" and argues that none can undermine his proposed standard. These include facts about AIs' having a different ontology, efficient cause, and final cause from the organisms that we consider moral patients. Danaher's strategy is to argue that none of these facts, by themselves or in combination with others, are relevant to whether an entity can be a moral patient. Regarding AIs' different ontology, for instance, he argues that whether an entity is made of organic or inorganic matter does not make any difference to whether it deserves moral consideration, and discriminating on that basis would be unjustified biological prejudice. Similarly, he argues that the fact that AIs are created through a different causal process or for a different purpose than biological organisms should not matter to how we treat them. It seems implausible, for instance, to think that babies born through a different causal process such as genetic enhancement deserve less moral consideration simply because their causal origin differs from other babies'. It is similarly implausible to think that whether we are designed or created by God to fulfill certain ends makes a difference to our claim to moral status (*ibid*, pp. 2031-2035).

Granted, characteristics like being made of organic matter or having evolved by natural selection seem irrelevant to moral patiency. There is no reason to suppose that it is impossible for AIs to be moral patients due to their different ontology or causal origin. This, however, does not mean there is nothing epistemically significant about these differences. Danaher assumes that our epistemic justification for ascribing patiency can only consist of behavioural evidence. As I argue below, however, this assumption overlooks the role of a special kind of evidence that plays a key role in our attributions of patiency in the paradigmatic cases, namely our first-personal knowledge of our own case. Once we recognize this role, we can see that AIs' differences from organisms do amount to epistemic defeaters.

To put it briefly, when it comes to many of the morally relevant capacities for which we do not have a fully worked-out philosophical account, we have first-personal knowledge of these capacities in our own case. We know, for instance, about our own conscious experiences, thoughts, and actions, and so we know we have the capacity for consciousness and the ability to form beliefs, desires, and intentions. Although our self-knowledge may not be infallible or complete, it does provide us with a form of epistemic access that is not available regarding entities other than ourselves. This knowledge of our own capacities is an indispensable part of our epistemic justification in attributing certain morally relevant capacities to other humans and animals that share our biological constitution and evolutionary history. This self-knowledge, however, cannot play the same epistemic role in the case of nonbiological intelligent machines whose constitution and causal origin does not resemble ours. That is why facts about the ontology and causal history of AIs amount to epistemic defeaters in their case.

In the context of the problem of other minds, similar arguments have been made by philosophers who defend the role of our self-knowledge in justifying our belief that other people or certain other animals have minds (Melnyk 1994; Sober 2000; Andreotta 2020). Melnyk (1994), for instance, argues that reference to our own case plays an indispensable evidential role in justifying the belief in other minds. He argues against the prevalent idea that what justifies our belief that other people have minds is simply an inference to the best explanation of their observed behaviour, i.e., the same kind of justification we have for believing in electrons or other theoretical entities.[30] Melnyk argues that the hypothesis that other people have minds can only be viewed as the *best* explanation of their behaviour if we take into account what we know about our own case. Without reference to our own case, there is hardly a reason to attribute consciousness or various qualitative mental states to any other entity. Consider, for instance, the hypothesis that someone exhibiting pain-behaviour is in fact in pain, i.e., having an unpleasant qualitative experience. An alternative hypothesis that explains the person's behaviour equally well is that the pain-behaviour is caused by the person's being in a complex physical state that plays the same causal role which the qualitative experience of pain is supposed to play, but does not involve any qualitative feeling. Why should we think the pain hypothesis is superior to the quale-free hypothesis? As Melnyk argues, it is only because of what we know about our own case that we can justifiably view the pain hypothesis as the best explanation of this person's pain-behaviour. To demonstrate this, Melnyk invites us to imagine a computer program that has mastered various forms of scientific reasoning, including inference to the best explanation. Presented with data describing the pain-behaviour of humans, such a computer can plausibly come up with some version of a quale-free hypothesis, but there would be no reason for it to even think to formulate the pain hypothesis (*ibid.*, p. 485). What enables *us* to form the pain hypothesis for explaining the behaviour of others is what we know about our own experience of pain and what causes us to feel it. Moreover, what makes the pain hypothesis a *better* hypothesis than the alternative is also this knowledge of our own pain and the conditions that cause it. It is more plausible to believe that what causes pain in me also causes pain in my conspecifics than to believe what happens in my case is somehow unique for no particular reason.

The significance of this reference to our own case is that the scope of the extrapolations we can justifiably make to other cases is limited. The guiding principle is that the same cause results in the same effect under the same conditions. In the case of other humans, since they have a similar biological constitution, brains, and nerve fibres, the most plausible explanation of their pain-behaviour seems to be that they too experience pain, just like I do. In the case of nonhuman living things, the basis for extrapolation is less obvious, but considerations of evolutionary continuity

[30] He argues specifically against Pargetter (1984), who defends such a view.

still provide a basis for explaining their behaviour in the same way. As Sober (2000) argues, there is reason to believe that organisms that are genealogically related to us share the proximate mechanisms that cause pain in us. This kind of common-cause reasoning, however, only works when there is in fact evidence for a common cause. Sober thus argues that the extrapolation does not work for imaginary extraterrestrials evolved independently of life on Earth, organisms whose shared behaviours are not homologies, or organisms known to deploy different neural machinery for exhibiting the shared behaviour (pp. 384-385). By the same token, when it comes to intelligent machines that are wired entirely differently from us and are constituted from different materials, there is no reason to believe the pain hypothesis explains their behaviour better than the quale-free hypothesis.

Moreover, in the case of many intelligent machines, an additional consideration tells against viewing the pain hypothesis as the best explanation of their behaviour. There is an alternative explanation of the behaviour of many AIs that is not available in the case of nonhuman organisms or imaginary extraterrestrials: they are intentionally designed so that their behaviour mimics ours. We can explain the *degree* to which they behave similarly to us simply by noting that their behaviour is a direct copy of ours. This makes the pain hypothesis even less necessary in this case. In the case of independently evolved organisms or extraterrestrials, a high degree of behavioural similarity can itself suggest that they may share the proximate mechanisms or evolutionary conditions that give rise to the capacity for feeling pain or other mental states in us. But in the case of AIs that are intentionally created—with different mechanisms and under different conditions—to behave exactly like us, no amount of behavioural similarity can be suggestive of a common cause.

The argument presented above focuses on the question of other minds and the justification for attributing consciousness to other entities. But the underlying reason for epistemically differentiating between biological organisms and AIs also applies to the question of moral patiency.[31] Our epistemic justification for considering ourselves as having the capacities that underlie moral patiency is not entirely or solely based on observing our own behaviour. And to the extent that our attribution of patiency to other humans and nonhuman animals relies on our beliefs

[31] In the case of the capacity for practical reason and action, for instance, some have argued that we have a kind of self-knowledge of our own intentions and reasons for action that is not simply based on observing our own behaviour (see Anscombe 1963; Velleman 1989; Setiya 2008).

about ourselves, the same epistemic ground is not available in the case of machines with a different constitution and causal history. This is not to reduce moral patiency or any of the capacities underlying this status to our own biological constitution or to deny that they *can* be realized via different physical materials and mechanisms. It is to say that the fact that *we* have certain characteristics can only support attributing them to creatures that share the material constitution and mechanisms that give rise to these characteristics in us.

## 8. Concluding Remarks

I began by asking whether AIs will someday become moral patients. I have argued that, while this is in principle possible, there is no good reason to believe it *will* in fact happen. As we have seen, most of the literature either focuses on the hypothetical case of an entirely imaginary form of AI, or starts from controversial assumptions about what grounds moral patiency. I have thus tried to offer an argument that, on one hand, can inform our present and future decisions about *actual* AIs, and on the other hand, does not presuppose any substantive account of moral patiency.

My argument relies on a minimal necessary condition that directly follows from the concept of moral patiency: the condition of having a good of one's own. I start from the plausible assumption that traditional artifacts like knives and cars do not meet this condition. I argue that although these artifacts are ascribed goals and functions, they are different from living organisms in that their teleological organization is derivative on our goals, intentions, and beliefs. I then turn to the case of AIs, and argue that they are not different from traditional artifacts in this respect. Focusing on the process that enables AIs to improve their performance and increase their intelligence—i.e., machine learning—I argue that nothing about this process should make us think that AIs will one day acquire a good of their own. In addition, I argue that considerations about the end result of this process should not make us think that AIs will meet the conditions of moral patiency either. Although the complex, intelligent, or even humanlike behaviour of future AIs might make it difficult to distinguish them from genuine cases of moral patiency, doing so will not be beyond our epistemic reach.

Essentially, I have tried to show that we have no more reason to be concerned about the moral status of today's AIs than we do for traditional artifacts. And we have no more reason to worry about the moral status of *future* AIs than we do for any other kind of thing that is subject to gradual or sudden change, be it a climate system, a geological formation, or a celestial object. Nothing about the nature of artificial intelligence makes it a particularly good candidate for acquiring moral patiency. And the common reasons given for this concern—from having a higher degree of autonomy and independence to exhibiting intelligent human-like behaviour—do not withstand closer scrutiny. Thus, the widespread concerns about the moral status of AIs are unwarranted.

## Acknowledgements

I would like to thank Antoine Dussault and John Basl for their thoughtful comments on previous drafts of this paper. I am also grateful to Justin Garson, Brandon Conley, Regina Rini, and Ben Winokur for helpful discussion.

## References

Anscombe, G. E. M. (1963) *Intention*, second edition. Oxford: Blackwell.

Anderson, E. (2004). Animal rights and the values of nonhuman life. In C. Sunstein & M. Nussbaum (Eds.), *Animal rights: Current debates and new directions* (pp. 277-298). Oxford: Oxford University Press.

Andreotta, A. (2021). The hard problem of AI rights. *AI & Society*, 36, 19-32. doi:10.1007/s00146-020-00997-x

Arbor, J. L. (1986). Animal chauvinism, plant-regarding ethics and the torture of trees. *Australasian Journal of Philosophy*, *64*(3), 335-339. doi: 10.1080/00048408612342551

Attfield, R. (1981). The good of trees. *Journal of Value Inquiry* 15(1), 35-54. doi: 10.1007/BF00136626

Basl, J., & Sandler, R. (2013). The good of nonsentient entities: Organisms, artifacts, and synthetic biology. *Studies in the History and Philosophy of Biological and Biomedical Science*, *44*, 697-705. doi: 10.1016/j.shpsc.2013.05.017

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology*, *27*(1), 79-96. doi: 10.1007/s13347-013-0122-y

Basl, J. (2019). *The death of the ethic of life*. Oxford University Press.

Bedau, M. (1991). Can biological teleology be naturalized? *The Journal of Philosophy*, 88, 647–655. doi: 10.5840/jphil1991881111

Bedau, M. (1992a). Where's the good in teleology? *Philosophy and Phenomenological Research*, 52, 781–806. doi: 10.2307/2107911

Bedau, M. (1992b). Goal-directed systems and the good. *The Monist*, 75(1), 34-51. doi: 10.5840/monist19927516

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines,* 22(2) 71-85. doi: 0.1007/s11023-012-9281-3

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

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1-12. doi: 10.1177/2053951715622512

Callicott, J. (1989). *In defense of the land ethic: Essays in environmental philosophy*, Suny Press.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and information technology,* 12(3), 209-221. doi: 10.1007/s10676-010-9235-5

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology*, *27*(1), 61-77. doi: doi.org/10.1007/s13347-013-0133-8

Crisp, R. (2006). *Reasons and the good*. Oxford: Clarendon Press.

Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*, *26*(4), 2023-2049. doi: 10.1007/s11948-019-00119-x

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books.

Feldman, F. (2004). *Pleasure and the good life*. Oxford: Clarendon Press.

Finnis, J. (1980). *Natural law and natural rights*. Oxford: Clarendon Press.

Fletcher, G. (2013). A fresh start for an objective list theory of well-being. *Utilitas* 25(2), 206–220. doi: 10.1017/S0953820812000453

Fulford, K. W. M. (1999). Nine variations and a coda on the theme of an evolutionary definition of dysfunction. *Journal of Abnormal Psychology*, 108, 412–420. doi: 10.1037//0021-843x.108.3.412

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, 62, 729-754. doi: 10.48550/arXiv.1705.08807

Griffin, J. (1986). *Well-being: Its meaning, measurement and moral importance*. Oxford: Clarendon Press.

Gordon, J. S. (2020). What do we owe to intelligent robots?. In J.S. Gordon (Ed.), *Smart Technologies and Fundamental Rights* (pp. 17-47). Brill Rodopi.

Gordon, J. S. (2021). Artificial moral and legal personhood. *AI & Society*, *6*(2), 457-471. doi: 10.1007/s00146-020-01063-2

Goodpaster, K. (1978). On being morally considerable. *The Journal of Philosophy*, 75(6), 308-325. doi:

Gunkel, D. J. (2018). The other question: Can and should robots have rights?. *Ethics and Information Technology*, *20*(2), 87-99. doi: 10.1007/s10676-017-9442-4

Gunkel, D. J. (2019). No Brainer: Why Consciousness is Neither a Necessary nor Sufficient Condition for AI Ethics. In *AAAI Spring Symposium: Towards Conscious AI Systems*.

Johnson, D. G., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology*, *20*(4), 291-301. doi: 10.1007/s10676-018-9481-5

Kamm F. (2007). *Intricate ethics: rights, responsibilities, and permissible harm*. New York: Oxford University Press.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*, M. Gregor (Trans. and Ed.), Cambridge: Cambridge University Press, 1998.

Korsgaard, C. (1996a). Kant's formula of humanity. In her *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press, pp. 106–132.

Korsgaard, C. (1996b). *The sources of normativity*. Cambridge: Cambridge University Press.

Korsgaard, C. (2013). The relational nature of the good. *Oxford studies in metaethics* 8(1), 1-26. doi: 10.1093/acprof:oso/9780199678044.003.0001

Korsgaard, C. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Kraemer, D. M. (2014). Revisiting recent etiological theories of functions. *Biology and Philosophy*, 29, 747–759. doi: 10.1007/s10539-014-9430-6

McLaughlin, P. (2001). *What Functions Explain: Functional Explanation and Self-Reproducing Systems*. Cambridge University Press.

McLaughlin, P. (2009). Functions and norms. In M. A. Perlman, F. Longy, & B. Preston (Eds.), *Functions in biological and artificial worlds: comparative philosophical perspectives* (pp. 93-102). MIT Press.

McShane, K. (2021). Against etiological function accounts of interests. *Synthese*, 198(4), 3499-3517. doi:

Melnyk, A. (1994). Inference to the best explanation and other minds. *Australasian Journal of Philosophy*, 72(4): 482-491. doi: 10.1080/00048409412346281

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science* 56(2), 288–302. doi:

Moosavi, P. (2019). From Biological Functions to Natural Goodness. *Philosophers' Imprint* (19), 1-20.

Mosakas, K. (2021). On the moral status of social robots: Considering the consciousness criterion. *AI & Society*, *36*(2), 429-443. doi: 10.1007/s00146-020-01002-1

Müller V., Bostrom, N. (2014). Future progress in artificial intelligence: a survey of expert opinion. In V. Müller (Ed.), *Fundamental issues of artificial intelligence,* (pp. 552-572). Berlin: Springer.

Müller, V. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, *23*(4), 579-587. doi: 10.1007/s10676-021-09596-w

Pargetter, R. (1984). The scientific inference to other minds. *Australasian Journal of Philosophy*, 62(2), 158–163. doi: 10.1080/00048408412341341

Neander, K. (1991). Functions as selected effects: the conceptual analyst's defense. *Philosophy of Science* (58), 168–84. doi: 10.1086/289610

Nicholson, D. (2013). Organisms ≠ machines. *Studies in the History and Philosophy of Biological and Biomedical Sciences* 44, 669–678. doi: 10.1016/j.shpsc.2013.05.014

Novelli, N. (2020). *On the granting of moral standing to artificial intelligence: a pragmatic, empirically-informed, desire-based approach*. Dissertation, University of Edinburgh.

Nozick, R. (1997). Do animals have rights?. In his *Socratic Puzzles*, Cambridge, MA: Harvard University Press, pp. 303–310.

O'Neill, O. (1998). Kant on duties regarding nonrational nature. *Proceedings of the Aristotelian Society Supplement* 72, 211–228. doi: 10.1111/1467-8349.00042

Regan, D. (2002). The value of rational nature. *Ethics*112, 267-291. doi: 10.1086/324235

Regan, T. (1976). Feinberg on what sorts of beings can have rights. *Southern Journal of Philosophy* 14(4), 485-498. doi: 10.1111/j.2041-6962.1976.tb01304.x

Roe, K., & Murphy, D. (2011). Function, dysfunction and adaptation? In: P. Adriaens, A. De Block (Ed.) *Maladapting minds: philosophy, psychiatry, and evolutionary theory* (pp. 216–237). Oxford University Press.

Sapontzis, S. (1987). *Morals, Reason, and Animals*. Philadelphia: Temple University Press.

Setiya, K. (2022). Intention. *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), E. N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/fall2022/entries/intention/>.

Schneider, S. (2019) *Artificial You,* Princeton. NJ: Princeton University Press.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy 39*, 98-119. doi: 10.1111/misp.12032

Singer, P. (1973). Animal liberation. In R. Garner, *Animal rights*, pp. (7-18). London: Palgrave Macmillan.

Singer, P. (1981). *The expanding circle*. Oxford: Clarendon Press.

Singer, P. (1993). *Practical Ethics*. Cambridge: Cambridge University Press.

Sober, E. (2000). Evolution and the problem of other minds. *The Journal of Philosophy, 97*(7), 365-386.

Sorabji, R. (1964). Function. *Philosophical Quarterly*, 14, 289–302. doi: 10.2307/2217769

Sullins, J. P. (2006). When is a robot a moral agent?. *The International Review of Information Ethics 6*, 23–30. doi: 10.29173/irie136

Sumner, L.W. (1996). *Welfare, Happiness, and Ethics*. Oxford and New York: Oxford University Press.

Taylor, P. W. (1986). *Respect for Nature: A Theory of Environmental Ethics.* Princeton: Princeton University Press.

Thompson, M. (2008). *Life and action: elementary structures of practice and practical thought*. Harvard University Press.

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *Washington post*, June 11, 2022 at 8:00 a.m. URL = <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>.

Tonkens, R. (2012). Out of character: on the creation of virtuous machines. *Ethics and Information Technology*, *14*(2), 137-149. doi: 10.1007/s10676-012-9290-1

Turing, A.M. (1950) Computing machinery and intelligence, *Mind*, 59, 433–460. doi:

Udell, D.B., & Schwitzgebel, E. (2021) Susan Schneider's Proposed Tests for AI Consciousness. *Journal of consciousness studies* 28.5-6, 121–144.

Varner, G. (1998). *In Nature's Interests? Interests, Animal Rights, and Environmental Ethics*. Oxford University Press.

Velleman, J. D. (1989). *Practical Reflection*. Princeton: Princeton University Press.

Walsh, D., & Ariew, A. (1996). A taxonomy of functions. *Canadian Journal of Philosophy*, 26(4), 493–514. doi: 10.1080/00455091.1996.10717464

Warren, M.A. (1997). *Moral status: Obligations to persons and other living things*. Oxford: Oxford University Press.

Weller, A. (2019). Transparency: Motivations and Challenges. In Samek, W., et al. (Eds.), *Explainable AI: interpreting, explaining and visualizing deep learning,* (pp. 23-40). Springer Nature.

Wood, A. (1998). Kant on duties regarding nonrational nature. *Proceedings of the Aristotelian Society Supplement* 72, 189–210. doi: 10.1111/1467-8349.00042

Woodfield, A. (1976). *Teleology*. Cambridge University Press.

Wright, L. (1973). Functions. *The Philosophical Review*, 82(2), 139-168. doi: 10.2307/2183766

Wright, L. (2013). Epilogue. In P. Huneman (Ed.), *Function: Selection and Mechanisms* (pp. 233-243). Springer.