Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09596-w

ORIGINAL PAPER

Is it time for robot rights? Moral status in artificial entities

Vincent C. Müller[1]

Accepted: 12 April 2021

Abstract
Some authors have recently suggested that it is time to consider rights for robots. These suggestions are based on the claim that the question of robot rights should not depend on a standard set of conditions for 'moral status'; but instead, the question is to be framed in a new way, by rejecting the is/ought distinction, making a relational turn, or assuming a methodological behaviourism. We try to clarify these suggestions and to show their highly problematic consequences. While we find the suggestions ultimately unmotivated, the discussion shows that our epistemic condition with respect to the moral status of others does raise problems, and that the human tendency to empathise with things that do not have moral status should be taken seriously—we suggest that it produces a "derived moral status". Finally, it turns out that there is typically no individual in real AI that could even be said to be the bearer of moral status. Overall, there is no reason to think that robot rights are an issue now.

Keywords AI · AMA · Artificial intelligence · Artificial moral agent · Ethical behaviourism · Ethics · Moral agent · Moral consideration · Moral patient · Moral status · Orchestration · Person · Relational turn · Rights · Robot

Introduction: rights, agents, patients and moral status

Some people in the field take the view that it should be seriously considered whether robots should be allocated rights now (Cappuccio et al., 2020; Coeckelbergh, 2012, 2018, 2020a; Danaher, 2020; Gunkel, 2018a, 2018b, 2020; Tavani, 2018; Turner, 2019). To be sure, the general issue who or what should have rights is of crucial importance—some of the worst atrocities in human history were and are based on wrongly claiming that certain humans or other animals do not have rights.

Before going into the details, we must briefly establish the state of the art and some terminology. James Moor distinguished four types of machine ethical agents: *ethical impact agents* (example: robot jockeys), *implicit ethical agents* (example: safe autopilot), *explicit ethical agents* (example: using formal methods to estimate utility), and *full ethical agents*. He concludes: "A full ethical agent can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent. We typically regard humans as having consciousness, intentionality, and free will." (Moor, 2006: 20; cf. Schlosser, 2019). Of course, 'free will' is a heavily contested notion, but it appears that the main criteria are intentional agency, alternative possibilities, and causal control (List, 2019; Mayr, 2011). In the philosophy of mind, a 'full agent' with 'free will' is often called a 'person' (Frankfurt, 1971) who is *responsible* for their actions. There is a long-standing and important discussion about the question which entities in the world are such responsible agents. Problematic cases include children, primates, people under drug influence, mentally handicapped people, future AI, angels, fictional characters, and social groups—the 'pessimists' think even typical human adults do not have such responsibility (Strawson, 2004).

In the discussions of robot rights, the less demanding notion of 'moral patient' plays a central role: As we saw, full ethical *agents* have rights and responsibilities while ethical *patients* only have rights, because harm to them matters. It appears that some entities are patients without being agents, e.g. animals that can feel pain but cannot make justified choices (Johnson & Verdicchio, 2018 says this should not lead us to treat robots this way). On the other hand, it is

✉ Vincent C. Müller
V.C.Muller@tue.nl
http://www.sophia.de

[1] Technical University Eindhoven (& University of Leeds & Alan Turing Institute), Eindhoven, Netherlands

Published online: 17 May 2021

generally assumed, with Kant, that moral agents or persons are always deserving of 'respect', and are thus moral patients as well. (This may well turn out to be problematic in the case of artificial moral agents.)

In this paper, we will use the term 'has moral status' to mean that an entity is a moral agent or a moral patient. (We do not differentiate between 'ethical' and 'moral', and we will not use the term 'rights' because this leads to further confusions.)

Apart from the terminology in this discussion, there are also authors who use the word 'agent' in a weaker sense, borrowing from the 'software agent' use in computer science, where matters of responsibility or even moral status will not arise. A characteristic remark is: "We conclude that there is substantial and important scope, particularly in Computer Ethics, for the concept of moral agent not necessarily exhibiting free will, mental states or responsibility." (Floridi & Saunders, 2004: 349). In a recent survey paper of the computing literature, the assumption is, quite simply, "The term AMA [artificial moral agent] will be used throughout the paper to refer to artificial agents capable of making ethical and moral decisions." (cf. Allen et al., 2000; Cervantes et al., 2020: 503). For the context of 'machine ethics' in AI ethics, see (Müller, 2020).

There are two issues that are mentioned (e.g. Gunkel, 2018b: 90, 93) in the discussion of 'rights for robots', but that have no bearing at all on the issue which entities have moral status, so we shall ignore them in the discussion: (1) Should we change our legal systems such that artificial systems or natural objects be treated as 'legal entities' or 'legal persons' with rights? (Bertolini & Aiello, 2018; Kurki, 2019: ch. 6; Stone, 1972) (2) Would it be ethical to make artificial systems that have phenomenal consciousness and thus moral status? Many people have their doubts (Bentley et al., 2018: 28f; Dennett, 2019; Schwitzgebel & Garza, 2015; van Wynsberghe & Robbins, 2019; Ziesche & Yampolskiy, 2019). Metzinger calls this a "negative synthetic phenomenology" or "The principle of avoiding artificial suffering" (Metzinger, 2013).

Reconstructions

Thinking otherwise

Gunkel's matrix

As explained above, the standard view is: If robots had some properties (especially sentience), then they would have moral status, but they do not have those properties now, so they do not have such status now. Gunkel has suggested (Gunkel, 2018a, 2018b) that we should not consider the question whether robots *do* have rights, but whether they *can* and *should* have rights. So, we are looking at two statements: "Robots can have rights" and "Robots should have rights", and their negations: a total of four combinations. (Surprisingly, there is no temporal dimension discussed here, so we take these to cover present and future.) He then tries to place positions in the current debate in the resulting matrix:

	Robots should not have rights	**Robots should have rights**
Robots cannot have rights	Instrumentalism; J. Searle	K. Darling
Robots can have rights	'Slaves'; J. Bryson	Properties; N. Bostrom, B. Goertzel; R. Sparrow

Out of the four theoretical options in the matrix, the two in the bottom row look like plausible options (can + should not; can + should), whereas in the top row, the first looks superfluous (cannot + should not), and the second contradictory (cannot + should).

Some authors, such as L. Floridi and A. Winfield are said to defend the orthodox belief that the notion of robot rights is "unthinkable", and thus to stand outside this matrix. Gunkel's book on the topic (2018a) is explicitly staged as "a provocative attempt to think about what has been previously regarded as unthinkable: whether and to what extent robots and other technological artifacts of our own making can and should have any claim to moral and legal standing" (http://machinequestion.org/robotrights/). But these authors do not say robot rights are 'unthinkable', they just say that current and foreseeable robots do not have the properties necessary for rights—not that robot rights are metaphysically impossible (i.e. contradictory) or epistemically impossible (i.e. impossible to think of). As Moor pointed out, "Even John Searle, a major critic of strong AI, doesn't argue that machines can't possess these features [of a full ethical agent]." (Moor, 2006: 21). A few authors have come close to such an impossibility claim, however, e.g. (Hakli & Mäkelä, 2019) argue that the history of robots (e.g. the fact that they are designed, and thus manipulated) prevents them from having responsibility.

Gunkel himself concludes, after many twists and turns, that the matrix does not work because the underlying assumption fails: There is no neat distinction between factual questions (Can robots have rights?) and normative questions (Should robots have rights?). He claims the is/ought distinction fails, and he suggests to "think otherwise"—thus the title of his paper (Gunkel, 2018b). Let us just note at this point that the difficulties with the matrix do not offer a reason to reject the is/ought distinction, which is a very large issue. But what is the argument, exactly, and what is it an argument for? C. Allen already

observed in his review of Gunkel's book (Gunkel, 2012) that the arguments are "often hard to discern" (Allen, 2013), so let us try a reconstruction.

Nasty reconstruction (1)

Here is an initial reconstruction of Gunkel's point. He looks at the standard argument against robot rights, which runs something like this:

1. Robots do not fulfil the requirements → Robots do not have rights
2. Robots do not fulfil the requirements
3. → Robots do not have rights

A simple modus ponens, starting from a necessary condition. But what is the response supposed to be? He rejects the 'properties view' in premise 1, so is this the argument?

(A) I deny premise 1
(B) I conclude 'not 3': Robots do have rights

Of course, concluding the negation of the conclusion is invalid, so that cannot be the point. In the best case Gunkel could conclude (B)' "I demand other reasons for conclusion 3." In that case, we need some support for claim (A) now. And what is that support for "thinking otherwise" supposed to be? Perhaps it is the 'relational turn'.

The relational turn

Background: humans caring about things (I)

Before continuing with this reconstruction, we need to remind ourselves of a fact about the relation between humans and artefacts. Humans quite easily forge an emotional bond with things in the world, and quite easily attribute some kind of as-if agency to things. Just remember how you respond to humanoid robots like *Pepper*, teddy bears or a gingerbread man—and even to objects that you hold dear even though they have little resemblance to living things, such as a violin, house, car or pencil. They seem to have value, somehow. In a famous study 80 years ago, Heider & Simmel showed participants "A motion picture which shows movements of three geometrical figures" [large triangle, small triangle, circle; see YouTube] and they found nearly all participants "interpreted the picture in terms of actions of animated beings, chiefly of persons" (Heider & Simmel, 1944: 243). This later lead to the 'attribution theory' as an account how humans explain the behaviour of others (Heider, 1956). These findings are consonant with results from human–robot interaction that Gunkel refers to, e.g. "We don't seem to care what their artificial intelligences 'know' or 'understand' of the human moments we might 'share' with them … the performance of connection seems connection enough" (Turkle, 2012: 9). The human tendency to attribution also raises the ethical question for human–robot interaction (HRI) whether we should build robots that exploit this tendency. – We shall have occasion to return to this matter later.

The relational turn

The suggestion in the 'relational turn' is that we should replace the orthodox account by an account that is based on the *relation* we humans have to the artefacts in question. This is achieved by "deliberately flipping the Humean script, considering not 'how ought may be derived from is' but rather 'how is is only able to be derived from ought.'" (Gunkel, 2018b: 95). This is supposed to be a radicalisation of Emmanuel Levinas' philosophy of encountering another person. In his book *Growing Moral Relations,* M. Coeckelbergh had said, programmatically: "Who is the architect, who constructs moral status? We humans do;" (Coeckelbergh, 2012: 7). In a later paper, he explains "'moral status' depends on moral status ascription and its conditions of possibility. … we should ask what kind of relations we want to have to them." and "This [human] moral subject is no longer a perceiver or observer but a doer, who actively relates to her environment." (Coeckelbergh, 2018: 197 and 208; cf. Coeckelbergh, 2020b: 47–62)—see also the criticism in (Gerdes, 2016). Gunkel summarises: "Consequently, the question of social and moral status does not necessarily depend on what the other is in its essence but on […] how we decide, in 'the face of the other' […], to respond." (Gunkel, 2018a: 97). We shall ignore the artful hedging with the words "social" and "necessarily" in that sentence.

There are several other authors who take this turn. (Dumouchel & Damiano, 2017) say in their book on social robotics, that we should think of robots as a "new social species" (xiii) because "… the traditional conception of emotions as discrete phenomena, as internal and private experiences, must be abandoned, and affect must be reconceived as continuous mechanism of interindividual coordination." (14). P. Gamez et al. propose to "investigate the 'machine question' by studying whether virtue or vice can be attributed to artificial intelligence; that is, are people willing to judge machines as possessing moral character?" (Gamez et al., 2020) (cf. Tavani, 2018). J. Bryson defines: "I intend to use … *moral patient* to mean "something a society deems itself responsible for preserving the wellbeing of," (Bryson, 2018: 16) and then she continues: "Human … moral systems … already attribute patiency to artefacts such as particular books, flags, or concepts" (Bryson, 2018: 16; cf. Wareham, 2020).

The idea that humans actively *make things important* was an important point that classical existentialism had stressed

already (Camus, 1942). Here, our human making this is taken to say we decide on what relations we want to have to "them", and this *makes* the moral status. If we decide to respond to something as a moral patient, then we *make it* into a patient, then it *is* a patient.

Nasty reconstruction (2)

Let us try a reconstruction of the argumentative structure here. For instances of "how we decide to respond" etc. that imply or generate moral patiency, we shall say "feel responsibility towards". So here is the argument:

1. I feel responsibility towards x → x has moral status
2. I feel responsibility towards robots
3. → Robots have moral status

This argument is valid (modus ponens) and premise 2 is often empirically true—as we indicated above, with attribution theory. But it is tempting to try a *reductio ad absurdum* by replacing premise 2 with another empirically true proposition: It so happens that I like nice pencils and I expect myself and others to treat them with care and respect. In other words, I feel responsibility towards pencils. So, here is vs. 2 of the argument:

1. I feel responsibility towards x → x has moral status
2. I feel responsibility towards pencils
3. → Pencils have moral status

This conclusion is pretty close to absurd. Premise 1, the core of the relational turn, is a version of *anything goes* that dissolves the question of moral patiency to a random act of will. Anything I happen to care about receives moral status. Indeed, I want you to respect my pencils, but should it follow that pencils have moral status? We shall return to other solutions below, under the heading of "derived moral status".

The absurdity remains if the "I" above is expanded to a "we", e.g. a society of pencil-worshippers. We, the members of that society can demand from the others that they should treat our pencils and customs with respect, but not that they should conclude that pencils have moral status, after all. (Just like you would not conclude that red-haired women lack moral status because some society thinks so.)

These consequences also remain if we add a little more caution than the "random" act of will:

1. I feel responsibility towards x and I want to feel responsibility towards x, upon reflection → x has moral status

Here we turned premise 1. into a 2nd order desire based on rational reflection, which is said to have the power to generate or to take away moral status.

We were not told, but the relational turn is a relativist account of moral status, with all the problems that come with that: no possibility to be right or wrong (to "respond" in the right or wrong way), no possibility of better or worse views, no possibility of moral progress, the ability to find out about moral status by just looking up what we/I made, etc. *Anything goes* now, including some conclusions that we now regard as sad low points in human moral history.

This is heavy baggage. Also, it weakens the proposal since standard moral relativism cannot criticise someone who replies "robots have no moral status for me/us". Perhaps it would be useful if the proponents of this theory could locate it in the standard landscape of relativism and anti-realism, both epistemologically and metaphysically. Are they perhaps of the meta-ethical view that declarative sentences about moral status (like "This robot is a moral patient.") should not be understood as making statements, as being true or false?

Danaher's "Behaviorism"

Exposition: performance and status

J. Danaher has suggested that for moral status ascription, we should neither rely on specific criteria for inner properties, nor on our tendency to respond (as in the relational turn), but rather on *observable performance*. He proposes:

(1) If a robot is *roughly performatively equivalent* to another entity whom [sic], it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.
(2) Robots can be roughly performatively equivalent to other entities whom, it is widely agreed, have significant moral status.
(3) Therefore, it can be right and proper to afford robots significant moral status. (Danaher, 2020: 2026)

This is presented as a version of methodological behaviourism, not ontological behaviourism, so it avoids an ontological thesis about *what there is*; instead it states *what we should talk about* in proper scientific method. To support premise (1), Danaher suggests a general view of 'ethical behaviourism', which "is the ethical equivalent of the Turing Test" (Danaher, 2020: 2028). The 'behaviour' here is said to include "measurable behaviour and brain phenomena, not inner mental states" (Danaher, 2020: 2028). In his conclusion, Danaher says: "If a robot looks and acts like a being to whom summarises moral status is afforded then it should be afforded the same moral status, irrespective of what it is made from or how it was designed/manufactured." (Danaher, 2020: 2047). So, if it walks like a duck, swims like a duck and quacks like a duck, it has the moral status of a duck!

This is a somewhat puzzling proposal, since we thought that behaviourism is dead … since its demise in the 1970ies when we realised that a) it does not capture what we mean by our mental terms, and b) it does not capture what we can do in cognitive science. Why drag it out from the graveyard?

Perhaps the solution to this puzzle is that the proposal is not quite behaviourism, and it is not quite about moral status, either: It does allow an investigation of 'inner states' ("brain phenomena"), and that sounds much more like standard cognitive science. States of the brain are not normally called 'behaviour' and looking into that 'black box' was specifically outlawed in the behaviourist programme. Of course, a cognitive scientist would use terms that can be grounded in scientific findings, not in folk psychology—and they would have to speak from a 3rd person perspective, but they would look at brain data and interpret that in terms of a functional, psychological, description, not just a description of the physical substrate. Danaher explains: "Ethical behaviourism states that a sufficient epistemic ground or warrant for believing that we have duties and responsibilities toward other entities (or that they have rights against us) can be found in their observable behavioural relations and reactions to us (and to the world around them)." (Danaher, 2020: 2028).

Yes, but the questions *what we have sufficient ground to believe* and *what is the case* are different questions. Even if I have sufficient ground to believe that an entity with a particular behaviour has moral status, it may still turn out that this entity does not in fact *have* moral status (similar points are made in (Nyholm, 2020, ch. 8.2). What causes ground or belief that something is a duck may be quite irrelevant to its moral status. An artefact like a robot may even be specifically designed to make us believe that it has moral status. So, it may walk like a duck, swim like a duck and quack like a duck, but not be a duck—and should not enjoy the moral status of a duck (which happens to be that or a moral patient). The initial argument set out above talks about what we should "afford" the robot, which is even more cautious—perhaps that means what we should 'provide it with'? In that case, robots do not have moral status, but we give it to them. Do we have such powers?

Nasty reconstruction (3)

Danaher comes to the same conclusion as the supporters of a 'relational turn' by taking 'what we normally attribute moral status to' as the criterion.

Of course, this works. We might take a statement like this:

(R) People believe that robots have moral status → robots have moral status

For 'believe' set your preferred version, e.g. 'behave as if', 'afford', 'relate to as if', 'normally attribute'. If one accepts a statement like (R), then the issue of moral status becomes a matter of our beliefs for any entity. This is especially true if we turn (R) into a biconditional:

(RR) People believe that robots have moral status ↔ robots have moral status

It is hard to see how someone might accept (R) but not (RR) because that would be to say that our belief can make robots gain moral status, but it cannot take away moral status from robots. As we saw above, *anything goes* if we accept that kind of statement.

There seems to be no issue of robot rights, just an issue of being tempted by versions of moral relativism. Why go down that road?

Diagnosis

Our All-Too human epistemic situation

Let us backtrack a bit and see whether there is some fire to all that smoke. Many of the points made in favour of the issue go back to our *epistemic* condition when we talk about moral status in entities other than ourselves. The criteria used in traditional accounts of moral status use features like phenomenal consciousness, free will and other mental states; and these have an especially problematic epistemic position. Danaher says: "The ethical behaviourist points out that our ability to ascertain the existence of each and every one of these metaphysical properties is ultimately dependent on some inference from a set of behavioural representations." (Danaher, 2020: 2029)—leaving out what we may know about inner workings or functional role. (Shevlin forthcoming) also comes to this conclusion, on the epistemic question alone. (Agar, 2019) says, more cautiously, there is an issue of how we should treat "machines that might have minds".

This is on the background of the traditional view that one's own experience is in principle inaccessible to other humans and the corresponding "other minds problem". The eminent philosopher of mind Thomas Nagel expresses the mainstream view when he says: "The only experiences you can actually have are your own: if you believe anything about the mental lives of others, it is on the basis of observing their physical construction and behavior."; "… your experience of tasting chocolate is locked inside your mind in a way that makes it unobservable to anyone else – even if he opens up your skull and looks inside the brain." Nagel concludes: "There seem to be two very different kinds of things going on in the world: the things that belong to physical reality, which many different people can observe from the outside, and those other things that belong to mental reality, which

each of us experiences from the inside in his own case." (cf. Nagel, 1974: 435; Nagel, 1987: 20f, 29f, 36).

This kind of discussion is surely in the back of our minds when we try to find ways to attribute moral status. This is not the place to go deeper these issues, but let it be taken as a given that we do not have a proper scientific method to test for the existence of phenomenal consciousness or free will—even if we assume that these are sensible concepts and that humans 'normally' have these features.

But what are we to conclude from such a situation? Does it follow that *anything goes,* in the style of principle RR), above? As we say, it would not help, anyway: If we humans can do as we please, I may attribute moral status while you do not (or *we* might, while *they* do not). Or does it follow from the epistemic situation that we should assume some wide precautionary principle, such as: "We cannot be absolutely certain that trees have no sentience, thus it is better to assume that they have sentience and treat them accordingly?" Surely not.

The real discussion is what we can know from behaviour and inner structure, and what is the best scientific theory that we can generate from that. The extended discussions about sentience in animals have some examples that are relevant to ours, e.g. "can fish feel pain?" (Allen & Trestman, 2017) or what are the conditions for artificial consciousness, e.g. (Seth, 2018) or in the *Journal of Artificial Intelligence and Consciousness*. This is analogous to the question what would be required to have a robot to feel pain, to mean what it says or to have 'machine consciousness'. These are real issues and we know what the methods for investigation are: They will involve conceptual clarification, theoretical models and scientific causal explanations that use behaviour and inner structure—but they do *not* include "how I decide to respond".

Note that we do not claim to have made a general case against moral status of robots, quite the contrary: We just tried to work out the assumptions in the proposals for attributing moral status to them *right now*—and find these problematic. One should attribute moral status to robots when they fulfil the criteria. This strikes us as less 'speciecist' than proposals that their moral status depends on 'how *we humans* decide to respond to *them* [sic!]' or 'what *we humans* decide to afford to *them*, given how they perform'. This is not to say that there are not *other* properties that may be response-dependent in a sense that can be explained (cf. Wheeler, 2020).

Last but not least, it is good to keep in mind that the criteria for moral status admit to a degree of vagueness, so we should not expect the set of 'objects with moral status' to have a sharp border (e. g. very young humans are not moral agents, but they can gradually grow to be agents). But even without sharp borders, there are objects that clearly fall into that set, and objects that do not.

Humans caring about things (II)—derived moral status

We said above that it seems absurd to claim moral status for pencils, but there is a bit more to this issue. It really would not be nice of you to step on my pencil, my teddy bear, the flag of my favourite football club or my religious symbol. These objects deserve what we might call 'consideration', meaning that behaviour towards them can be judged as morally right or wrong. Is that because they have moral status as moral patients? No. It is because *I* have moral status and thus deserve consideration. Doing something to these objects does something morally relevant *to me*. This would apply particularly to an object that represents me, in some way (e.g. my avatar) or have a special symbolic social role. If we want, we might call this a 'moral status' of objects, but this is only a *derived* moral status—rather like Searle's 'derived intentionality' of symbols (Searle, 2004: 19f) and 'derived value' of social constructs like money (Searle, 1995: 31ff). Ph. Brey had already suggested 13 years ago (against Floridi & Saunders) that "many (but not all) inanimate things in the world deserve moral respect, not because of intrinsic value, but because of their (potential) extrinsic, instrumental or emotional value for persons." (Brey, 2008: 109). M. Coeckelbergh also seems to take this line into account, recently (Coeckelbergh, 2020c).

It has been pointed out many times that artefacts like robots are more than just tools; they mediate between the user and the world. This has led some authors to propose an extension of moral agency to technological artefacts (a survey in Noorman, 2020). The outcome of these discussions appears to be just a conceptual clarification of 'moral agent', however.

Further to the consideration for objects, we will also often make judgments about *people* on the basis of their behaviour towards objects, even if these objects do not 'deserve consideration'. Deliberately stepping on someone else's teddy bear is problematic (a lack of respect for the other person), but so is stepping on one's own teddy bear (it shows a lack of self-respect). But even if there are no other persons involved at all, if there is no derived moral status, we tend to judge people's behaviour, and we might even assume that behaviour towards objects without moral status influences behaviour towards entities *with* moral status, e.g. animals or humans. So, there is a real question of how we should behave towards objects without moral status, including robots; perhaps sex robots are a good example of this (Nyholm, 2020, ch. 2; Whitby, 2008). As Dumouchel and Damiano rightly say, "how we live with robots … reflects our own moral character" (2017: xiv)—though surely this is not a reason to treat robots as moral patients, either (pace Cappuccio et al. 2020).

Beyond agency: orchestration

As an outlook to where the case of AI might lead us, we would suggest looking at some work on embodied cognition. Work on the notion that the morphology of the body 'computes' cognitive properties found that the dynamic orchestration of behaviour involves a complex interaction of computational and morphological features. In an earlier paper, we had asked "Perhaps we should abandon the old image of the central control in computers, but also in humans?" (Müller, 2007: 112) and later we concluded: "The question for robot design and cognitive science is not whether computation is offloaded to the body, but to what extent the body facilitates cognition and control—how it contributes to the overall orchestration of intelligent behavior." (Müller & Hoffmann, 2017: 1). In that view, intelligent behaviour is less the result of an agent acting in some way, but of a complex interaction of body, physical environment, social environment, cognition and interaction over time.

So, even in the design of robots, individuated agency plays a minor role; what matters is the orchestration of intelligent behaviour in a context. Of course, this is not to claim that there is no agency in the natural world, but rather that the aim in current AI technology is to produce a system that works, not an artificially intelligent individual. Whether the traditional and often metaphorical talk about 'agents' is still useful is debated in AI, quite independently from the acceptance or rejection of an embodied approach (Dignum & Dignum, 2020). That debate is worth continuing (in a different paper). Note how the talk about 'robot rights' glosses over the fact that artificial systems often have no particular body, or any other 'identity' or 'self' that could be said to 'have' a moral status. Criteria for identity are easier to come by for physical robots than for most AI systems. The question of moral status of agents and patients in AI may not even present itself now. It looks like a fiction for philosophers.

Conclusions

We started our hermeneutic circle on robot rights from "Introduction: Rights, Agents, Patients and Moral Status" the orthodox account of necessary conditions for moral status and then looked at "Thinking Otherwise" and "The Relational Turn", and an attempt "Danaher's "Behaviorism" to revive behaviourism, all of which ended up in the same relativist impasse. We suggested "Diagnosis" to return to the orthodox account, but to take home some lessons from this discussion:

(a) Humans tend to empathise with various things in the world, especially those that remind us of natural agents
(b) Our epistemic status with respect to 'other minds' is a problem in allocating moral status
(c) There are objects that do not have moral status but deserve consideration due to their relation with an agent/patient; they have a *derived* moral status
(d) Allowing (non-derived) moral status to be dependent on a human decision or attitude is philosophically unfounded and dangerous
(e) How humans behave towards objects that do not have moral status can be relevant for how we judge those humans
(f) Present AI does not aim for individual agents or patients but rather for orchestration of intelligent behaviour in a context

So, the attempts to re-define the issue of moral status for robots or AI systems made very substantial philosophical assumptions, but despite these they fail in their attempt to establish that 'robot rights' are an issue that is relevant for current or near-term systems. The question whether present-day robots have moral status is settled: They do not. But the discussion raises interesting issues about epistemology, about what human behaviour towards robots says about us humans, and about our all-too-human tendency to see individuals with moral status everywhere.

Acknowledgements I am grateful to audiences in Eindhoven, Montréal, Paris and Vilnius, as well as to David Gunkel for a discussion in Vilnius. My thanks for very useful comments to Sven Nyholm, Merel Noorman, Fabio Tollon and to my PhD students Gabriela Arriagada-Bruneau, Michael Cannon and Zach Gudmundsen.

References

Agar, N. (2019). How to treat machines that might have minds. *Philosophy & Technology, 33*, 269–282

Allen, C (2013). Review of D. J. Gunkel, The machine question: Critical perspectives on AI, robots, and ethics, MIT Press, 2012', *Notre Dame Philosophical Reviews,* February 13, 2013. https://ndpr.nd.edu/news/the-machine-question-critical-perspectives-on-airobots-and-ethics/.

Allen, C., & Trestman, M. (2017). Animal consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. (Winter 2017 ed.,) CSLI. https://plato.stanford.edu/archives/win2017/entries/consciousness-animal/

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261

Bentley, P. J., Brundage, M., Häggström, O., & Metzinger, T. (2018). Should we fear artificial intelligence? In-depth analysis. *European Parliamentary Research Service, Scientific Foresight Unit (STOA),* (PE 614.547), 1–40. Retrieved March, 2018, from <http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA%282018%29614547_EN.pdf. Accessed 11 Nov 2020.

Bertolini, A., & Aiello, G. (2018). Robot companions: A legal and ethical analysis. *The Information Society, 34*(3), 130–140

Brey, P. (2008). Do we have moral duties towards information objects? *Ethics and Information Technology, 10*, 109–114

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15–26

Camus, A. (1942). *Le mythe de Sisyphe*. Gallimard.

Cappuccio, M. L., Peeters, A., & McDonald, W. (2020). Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology, 33*(1), 9–31

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics, 26*, 501–532

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

Coeckelbergh, M. (2018). What do we mean by a relational ethics? Growing a relational approach to the moral standing of plants, robots and other non-humans. In A. Kallhoff, M. D. Paola, & M. Schörgenhumer (Eds.), *Plant ethics.* (pp. 110–121). Routledge.

Coeckelbergh, M. (2020a). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Coeckelbergh, M. (2020b). *AI ethics*. MIT Press.

Coeckelbergh, M. (2020c). Should we treat Teddy Bear 2.0 as a Kantian dog? Four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. *Minds and Machines, 30*, 1

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26*, 2023–2049

Dennett, D. C. (2019). Will AI achieve consciousness? Wrong question. *Wired.* Retrieved February 19, 2019, from https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/.

Dignum, V., & Dignum, F. (2020). Agents are dead. Long live agents! In N. Yorke-Smith, B. An, A. E. F. Seghrouchni, & G. Sukthankar (Eds.), *Proc. of the 19th International Conference on autonomous agents and multi agent systems AAMAS 2020.*IFAAMAS.

Dumouchel, P., & Damiano, T. (2017). *Living with robots, trans. Malcolm DeBevoise*. Harvard University Press.

Floridi, L., & Saunders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349–379

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 1*, 5–20

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society, 35*(4), 795–809

Gerdes, A. (2016). The issue of moral consideration in robot ethics. *SIGCAS Comput. Soc., 45*(3), 274–279

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

Gunkel, D. J. (2018a). *Robot rights*. MIT Press.

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

Gunkel, D. J. (2020). The rights of (killer) robots. In S. C. Roach & A. E. Eckert (Eds.), *Moral responsibility in 21st century warfare: Just war theory and the ethical challenges of autonomous weapon systems.* (pp. 1–21). CUNY.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275

Heider, F. (1956). *The psychology of interpersonal relations*. John Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*, 243–259

Johnson, D. G., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology, 20*(4), 291–301

Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press.

List, C. (2019). *Why free will is real*. Harvard University Press.

Mayr, E. (2011). *Understanding human agency*. Oxford University Press.

Metzinger, T. (2013). Two principles for robot ethics. In J-P. Günther & E. Hilgendorf (Eds.), *Robotik und Gesetzgebung*. Nomos. https://www.nomos-elibrary.de/10.5771/9783845242200/robotik-und-gesetzgebung.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21

Müller, V. C. (2007). Is there a future for AI without representation? *Minds and Machines, 17*(1), 101–115

Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Eds.), *Stanford Encyclopedia of Philosophy*. (pp. 1–70). CSLI Stanford University. <https://plato.stanford.edu/entries/ethics-ai/>

Müller, V. C., & Hoffmann, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artificial Life, 23*(1), 1–24

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*(4), 435–450

Nagel, T. (1987). *What does it all mean? A very short introduction to philosophy*. Oxford University Press.

Noorman, M. (2020). Computing and moral responsibility. In E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy*. CSLI Stanford University. <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility>

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.

Schlosser, M. (2019). Agency. In E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy*. CSLI Stanford University. <https://plato.stanford.edu/archives/win2019/entries/agency/>

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy, 39*, 98–119

Searle, J. R. (1995). *The construction of social reality*. Free Press.

Searle, J. R. (2004). *Mind: A brief introduction*. Oxford University Press.

Seth, A. K. (2018). Consciousness: The last 50 years (and the next). *Brain and Neuroscience Advances, 2*, 1–6

Shevlin, H. (forthcoming), How could we know when a robot was a moral patient?. *Cambridge Quarterly of Healthcare Ethics*

Stone, C. D. (1972). Should trees have standing-toward legal rights for natural objects. *Southern California Law Review, 2*, 450–501

Strawson, G. (2004), 'Free will', *Routledge Encyclopedia of Philosophy* (updated 2011). Retrieved May, 2005, from https://www.rep.routledge.com/articles/thematic/free-will/v-1. Accessed 11 Nov 2020.

Tavani, H. T. (2018). Can social robots qualify for moral consideration? Reframing the question about robot rights. *Information, 9*(73), 1–16

Turkle, S. (2012). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Turner, J. (2019). *Robot rules: Regulating artificial intelligence*. Springer.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*(3), 719–735

Wareham, C. S. (2020). Artificial intelligence and African conceptions of personhood. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-020-09541-3

Wheeler, M. (2020). Deceptive appearances: The turing test, response-dependence, and intelligence as an emotional concept. *Minds and Machines, 30*, 1–20

Whitby, B. (2008). Sometimes it's hard to be a robot. A call for action on the ethics of abusing artificial agents. *Interacting with Computers, 20*(3), 326–333

Ziesche, S., & Yampolskiy, R. V. (2019). Do no harm policy for minds in other substrates. *Journal of Evolution and Technology, 29*(2), 1–11