Ethics and Information Technology (2021) 23:147–155
https://doi.org/10.1007/s10676-020-09540-4

ORIGINAL PAPER

The artificial view: toward a non-anthropocentric account of moral patiency

Fabio Tollon[1]

Published online: 1 June 2020

Abstract
In this paper I provide an exposition and critique of the Organic View of Ethical Status, as outlined by Torrance (2008). A key presupposition of this view is that only moral patients can be moral agents. It is claimed that because artificial agents lack sentience, they cannot be proper subjects of moral concern (i.e. moral patients). This account of moral standing in principle excludes machines from participating in our moral universe. I will argue that the Organic View operationalises anthropocentric intuitions regarding sentience ascription, and by extension how we identify moral patients. The main difference between the argument I provide here and traditional arguments surrounding moral attributability is that I do not necessarily defend the view that internal states ground our ascriptions of moral patiency. This is in contrast to views such as those defended by Singer (1975, 2011) and Torrance (2008), where concepts such as sentience play starring roles. I will raise both conceptual and epistemic issues with regards to this sense of sentience. While this does not preclude the usage of sentience outright, it suggests that we should be more careful in our usage of internal mental states to ground our moral ascriptions. Following from this I suggest other avenues for further exploration into machine moral patiency which may not have the same shortcomings as the Organic View.

Keywords Machine moral patiency · Sentience · Anthropocentrism · Intentional stance · Organic view of ethical status

Introduction

When evaluating moral situations, we tend to think in terms of giving moral stakeholders their due: giving them what they *deserve* based either on how they have behaved or whether they have been harmed. There arises, firstly, the question of whether an entity is misbehaving *intentionally*, in the common-sense usage of the term ("on purpose"), and whether it could in some sense be *responsible* for its behaviour, and hence possibly morally responsible. This is a question of moral *agency.* Conversely, a second question may arise of whether, if we were to harm the entity, we would be doing it a *moral harm*. In other words, do we owe it certain moral *obligations*? This is a question of moral *patiency*. These two questions can be viewed as fundamental to all moral philosophy: *who* or *what* is deserving of moral concern, and *who* or *what* can be said to be (morally) responsible for their actions (Gunkel 2012, p. 1). Moral patients are the class of entities that can in principle qualify as *receivers* of moral action, whereas moral agents are the class of entities that can in principle qualify as *sources* of moral action (Floridi and Sanders 2004, pp. 349–350).

On the one hand, trends in contemporary macro-ethics have been geared toward expanding the boundaries of moral consideration by focusing on the nature of who or what should count as a moral patient. This ascription of moral patiency is independent of whether the entity in question is a moral agent or not (Floridi and Sanders 2004).[1] However, while not all moral patients are moral agents, it is standardly supposed that all moral agents are moral patients (see Floridi and Sanders 2004; Torrance 2008). On the other hand, the

✉ Fabio Tollon
fabiotollon@gmail.com

[1] Philosophy Department, Stellenbosch University, Stellenbosch, Western Cape, South Africa

[1] A patient-orientated approach to ethics is not concerned with the perpetrator of a specific action, but rather attempts to zero in on the *victim* or receiver of the action (Floridi, 1999). This type of approach to ethics is considered non-standard and has been incredibly influential in both the "animal liberation" movement and "deep ecology" approaches to environmentalism (see Leopold, 1948; Naess, 1973; Singer, 1975, 2011). Both place an emphasis on the *victims* of moral harms; in the case of animal liberation, the harm we do to animals, and in the case of deep ecology the harm we do to the environment.

emergence of artificially intelligent systems, properly conceptualised as artificial agents[2] (AAs), may complicate many presuppositions of what counts as a moral action. These systems may come to undermine the standard assumption above by performing actions which, while independent of human control, might still be subject to moral assessment (see Sparrow 2007; Johansson 2010; cf. Johnson and Noorman 2014; Johnson 2015). While the latter question is deserving of (and has received) considerable philosophical attention, my focus in this paper will not be concerned with moral agency directly. Instead, I will assume the validity of the conceptual relationship between agents and patients which claims that all moral agents are moral patients. It is with the aforementioned in mind that any investigation into moral agency must first address the question of moral patiency. This conceptual point stresses the importance of the discussion in this paper, as the implications of this approach for machines are clear: if machines cannot be considered moral patients, then they cannot be moral agents either. The stakes in this debate are quite high. If we were to conclude that no computationally-based systems can ever be fitting subjects of moral concern, then our treatment of them need not follow any moral contours. Our treating them and their needs as morally subordinate to our own would not be problematic, as we would owe them no moral obligations. However, if it turns out that we were wrong to treat these systems as "mere machines", then we would find ourselves guilty of harming an entirely new class of moral patient, and unjustifiably excluding them from our moral universe. It is therefore important that we take the "machine question"[3] seriously (Gunkel 2012, p. 5).

The organic view of ethical status

In order to address the question of machine moral patiency I will provide an exposition and critique of the Organic View of Ethical Status (hereafter simply the "Organic View"), as it is articulated by Steve Torrance (2008). The Organic View makes an important contribution to the philosophical debate on moral status. Torrance's exposition of the Organic View brings together many characteristics that make a consistent appearance in the literature on machine moral agency and patiency. These are questions of sentience, intentionality, and the conceptual relationship between moral agents and moral patients (see Floridi and Sanders 2004; Johnson and Miller 2008; Himma 2009; Sullins 2011; Johnson and Noorman 2014).

The Organic View raises pertinent ethical questions, specifically, whether the expansion of our "mental" universe to include machines also necessitates an expansion of our moral universe to include them (Torrance 2013, p. 399). In order to make his case, Torrance centres his discussion around two factors which feature prominently in the Organic View: firstly, he claims that *sentience*[4] (or phenomenal consciousness) is a key factor in the type of rationality moral entities exhibit, and, secondly, that *biological constitution* is of fundamental moral significance (2008, p. 505). This paper focuses on the first of these claims, and while Torrance does not explicitly endorse the Organic View, he does seem to harbour a favourable disposition toward it. While he is willing to concede that it may well be wrong (or at the very least in need of further qualification) (Torrance 2008, p. 505), I will endeavour to show, in line with the work of Mark Coeckelbergh, that the Organic View succumbs to issues of *justification* in terms of moral consideration (2010a,2014; b). The content of this paper is therefore broadly in line with Coeckelbergh's project: it takes seriously the expansion of our moral universe in a way that does not rely only on ontological features of the entity in question (2010, p. 212). I will show how the Organic View gives us a philosophically interesting way in which to view the moral status of artificial systems, but that it nonetheless still falls victim to the issues raised by Coeckelbergh (2010a, b 2014). In order to make my argument I first put forward the case made by Torrance (2008, p. 503) that AAs do not have "empathic rationality", with the implication that machines, unless they can be designated as "sentient", cannot be proper subjects of moral concern. From this, I then show how the sense of sentience Torrance operationalises in his account is flawed due to both conceptual and epistemic shortcomings.

Empathic rationality

In this section I deal with a specific (but essential) claim of the Organic View: "Only beings which are capable of sentient feeling or phenomenal awareness could be genuine subjects of either moral concern or moral appraisal" (ibid., p. 503). The reason for focusing on this aspect of the Organic View is that, if found wanting, it would undermine the entire

[2] An artificial agent is artificial in the sense that it has been manufactured by intentional agents (us) out of pre-existing materials, which are external to the manufacturers themselves (Himma, 2009). It is an agent in the sense that it is capable of performing actions (Floridi and Sanders, 2004: 349). An easy example of such an artificial agent would be a cellphone, as it is manufactured by humans and can perform actions, such as basic arithmetic functions or responding to queries via online searches.

[3] Gunkel (2012: 5) considers the "machine question" to be the flip side of the "animal question": both concern the moral standing of non-human entities.

[4] Sentience can be understood as the capacity for an entity to have phenomenal/subjective/qualitative states of experience (Bostrom and Yudkowsky, 2011: 7).

argument. The criterion of sentience is what grounds Torrance's conception of moral patiency, and so if it can be found wanting it would be a serious threat to the validity of the argument. This will become clear as my critique develops.

Torrance begins his argument by asking us to imagine an AA that has a certain minimum level of rationality and has the cognitive ability to recognise that certain beings have sentient states, and thus moral interests (Torrance 2008, p. 510). Moreover, the AA can reason about the effects that different courses of action may have on these sentient creatures. Yet, this type of agent does not have the capacity to *feel* moral concern (ibid.). Such agents, due to their ability to *cognitively* apprehend and interpret the behavioural cues of other entities, and to infer from these that the entity in question could be undergoing a moral harm, etc., *may* be thought of as being fitting subjects of moral appraisal (ibid.). Due to their ability to cognitively apprehend and reason about moral situations, these entities could use this ability to guide their actions – these then being subject to moral evaluation. In other words, by fulfilling certain rationality criteria (which other non-human entities do not), one might think it reasonable to extend the ascription of moral agency to these entities; even if they are not sentient in the same way that human beings are (ibid.).

Nevertheless, the problem with this view, according to Torrance (ibid.), lies in assuming that the type of rationality required for moral agency is simply cognitive or intellectual, as this would provide us with an anaemic account of moral standing. Torrance suggests that the kind of rationality required for an entity to legitimately be given the status of moral agent may turn out to be different from the kind that could be achieved by an AI system. He argues that the type of rationality traditionally associated with our own full moral status (as humans) is closely associated with our sentient nature (in other words, our capacity for *affect*) (ibid.). Thus the claim is that being a moral agent requires (human) sentience (or affect) (ibid.). The argument goes as follows: our kind of rationality involves the capacity for a kind of affective or empathetic identification with the experiential states of others, where such identification is integrally available to the agent as an essential component in its moral decision-making procedures (ibid.). Torrance (ibid.: 516) calls this kind of rationality *empathic rationality* and contrasts it with the purely *cognitive* or *intellectual rationality*, which might be attributable to intelligent, computationally-based AAs. While we expect information-processing systems to make decisions in a purely mechanistic way, Torrance claims that we have different standards when it comes to our moral decision-making procedures, as we expect human beings to factor the potential experiential consequences of their actions into their moral reasoning (ibid., p. 511). Significantly, he claims that entities which are only capable of intellectual rationality would not have a "real" or "true" means of evaluating the experiential states of others. Such an entity could simply not understand how its actions might affect others.

Thus, Torrance's argument is that moral decision making requires the capacity for "engaged empathic rational reflection" (ibid., p. 511), which requires the ability to identify with the experiential states of others. Any rational agent that is not also sentient (in a manner equivalent to the type of sentience achievable by biological organisms) would not have this empathic ability, since a precondition for a "true" understanding of experiential states is that one is able to have these states oneself. Since only entities capable of being "ethical consumers" can have this type of empathic rationality, other types of agents are precluded from being subject to moral evaluation, as without the ability to take a "moral point of view", it would be a mistake to then evaluate actions undertaken by such agents using moral criteria (ibid., p. 499). The Organic View suggests, then, that we should conclude that entities lacking a specific type of sentience cannot be moral agents.

Problems with the organic view of ethical status

The first ambiguity that needs to be addressed is the vague way in which internal, experiential states are operationalised in Torrance's articulation of the Organic View. Only organisms capable of having some kind of "qualitative experience" of pain (or any other such experiential state) will qualify as moral patients (and by extension, according to the Organic View, as moral agents).[5] Moreover, as Torrance (2014) is a realist about mental states, he claims that there is an *objective* answer when we ask the question as to an entity's psychological state.[6] This realism about mental states works to buttress his views regarding our moral ascriptions to artificial entities: Torrance's specific form of realism claims that even if there were no functional or cognitive difference between an artificial and biological system, there

[5] For the sake of argument, I focus here on the experience of pain, but logically it would be possible to subject any type of internal mental state to the same type of analysis. Any theory which posits an "experience of *X*" claim must eventually answer to the question of *who* or *what* (i.e. what *type of mind*) is *experiencing*, or capable of experiencing, *X*.

[6] Torrance does not believe that functionalist accounts of mind fully capture the qualitative aspects of experience. He thus believes in the metaphysical possibility of "philosophical zombies"; humans which look and behave indistinguishably from us but lack phenomenal conscious states of experience (Torrance, 2008). This is a thorny philosophical issue in its own right, but I will not go into further detail here.

would still be a *phenomenal*[7] difference (ibid., p. 13).[8] This phenomenal difference is of fundamental moral significance for Torrance given that he claims some biological form of sentience is a prerequisite for moral patiency. In what follows, I will, firstly, bring to light conceptual ambiguities inherent to the Organic View, and secondly, discuss how the epistemic distinction between the mere "appearance" of something and the "real thing" operationalised in the Organic View is a problematic one.

Conceptual Issues

To see the ambiguity more clearly, an example put forward by Daniel Dennett (1996) offers a wonderful (albeit grisly) illustration of this. Dennett asks us to imagine that:

> A man's arm has been cut off in a terrible accident, but the surgeons think they can reattach it. While it is lying there, still soft and warm, on the operating table, does it feel pain? A silly suggestion you reply; it takes a mind to feel pain, and as long as the arm is not attached to a body with a mind, whatever you do to the arm can't cause suffering in any mind. (ibid., pp. 16–17)

Our intuition is that, although it might be possible to argue that the detached arm on the table may be capable of adverse nerve stimulus (i.e. pain), without being attached to some kind of mind this pain can never constitute suffering. The *experience* of pain is equivalent to suffering, and without an *experiencer* pain in itself can be of no moral significance (Gunkel 2012, p. 115). At this point a defender of the Organic View can agree, as this seems to be the exact point that they are arguing for, as only *genuinely* sentient creatures would be deserving of moral concern. Such sentient creatures are the equivalent of an "experiencer of pain" in the example above, in that they are the "experiencers of moral violation"; however, in what follows I will argue that this is a problematic stance to adopt.

While it might be reasonable to attribute the status of moral patient to certain classes of sentient animals, as we go further down the phylogenetic tree, and as creatures differ from us in their external appearance, we tend to be less likely to attribute the requisite kind of sentience to them. Torrance acknowledges this issue but remains neutral on whether this is a strength or a weakness of the Organic View (2008, p. 515). I believe it is a weakness of the position, as one of the main reasons for the issue arising in the first place is Torrance's emphasis on biological sentience as a key ontological property of entities deserving of moral concern. As Gunkel points out, there seems to be an "irreducible terminological slippage associated with this concept" (2012, p. 115). Moreover, "suffering easily becomes conflated with and a surrogate for consciousness and mind" (Gunkel 2012, p. 115). We are inclined to view other *hominids* as sentient, but most would not award this same ascription to other creatures which perhaps have more "basic" minds, such as molluscs. We tend to think of them as analogous to the arm on the table: capable, perhaps, of adverse nerve stimulus, but not *sentient* to the required degree, not capable of *experiencing* pain. Moreover, the Organic View itself does not give us a clear criterion for sentience (of the requisite kind), and so we have to rely on our intuitions to determine which kinds of creatures are moral patients, and these intuitions are geared towards including those entities that look like us and excluding those that look less like us.

These intuitions do not necessarily track "actual" sentience, and so the criterion of sentience does not help us, in practice, to identify moral patients. Gunkel (2012) makes a similar point when discussing the various issues surrounding our identification of suffering and pain in animals. He states that while it *seems* our intuitive ascriptions make sense, we still do not have a settled answer to the question for what distinguishes pain from suffering (Gunkel 2012, p. 115). To see this more clearly consider the example of fish, more specifically, fish cognition. Our *perception* of an animal's intelligence is often a key criterion (although not the only one) for whether we consider them to be sentient or not, and fish are rarely considered to be intelligent or phenomenally sentient in a manner akin to humans or even mammals. Moreover, fish are very rarely (if ever) accorded the same type of moral concern as are warm-blooded, non-human animals. Standard reasons given for such claims is that fish lack the requisite neural complexity in order to have the right kind of "experience". Such endothermism[9] (in the case of fish, specifically) stems from a disjunction between the public perception of fish intelligence and scientific reality (Brown 2015). There is ample scientific evidence supporting the conclusion that "fish perception and cognitive abilities often match or exceed other vertebrates" (*ibid.*). For example, fish are capable of tool use and display evidence of complex social organisation and interaction (such

[7] Phenomenal in the sense of having the capacity for conscious awareness. When applied to his argument for moral status, however, Torrance does not require that the entity in question be self-aware, only sentient (2008: 503).

[8] My own view is that there is in fact no difference between what can be "functionally" known about the mind and "phenomenal" aspects of mind: the phenomenal is just a special case of the functional, and in this way, there is no "hard problem" of consciousness. See Chalmers (1996) for a defense of the hard problem, and Cohen and Dennett (2011) for a substantive critique.

[9] That is, unfair moral discrimination based on the temperature of an entity's blood.

as signs of cooperation and reconciliation) (*ibid.*). The point here is not to outline all of the ways in which fish cognition may be measured. Rather, the key issue is that if we use our traditional metrics of intelligence when it comes to animals (such as tool use and social organisation), then we are forced to conclude that fish are on par with (and at times exceed) other "sentient" vertebrates in these criteria. The next question, then, would be whether, following from the fact that fish exhibit "intelligent" behaviour, they are also phenomenally sentient and hence capable of similar kinds of suffering? Our intuitions surrounding fish sentience and their capacity to feel and suffer seem to be biased away from accepting them as sentient "enough" to merit moral concern. It seems that we struggle to empathise with fish as.

> [w]e cannot hear them vocalise, and they lack recognisable facial expressions both of which are primary cues for human empathy. Because we are not familiar with them, we do not notice behavioural signs indicative of poor welfare. (*ibid.*)

This implies that a proper, scientific construal of fish behaviour would support the conclusion that fish have relatively complex cognitive capacities, are capable of suffering, and are therefore sentient in a manner similar to creatures that are accorded moral concern (ibid.). To bring this back to the Organic View, the issue that the example above was meant to highlight is that how we go about identifying moral patients should not be guided by concepts with have intractable conceptual slippage associated with their usage.

Applying the discussion above to the question of whether an artificial system could, in principle, be the subject of moral concern, highlights the potential for moral harm in the future. In the same way that we have biases that cause us to accord a lesser moral status to non-human entities that do not sufficiently look like us, we may be biased against machines based on their unfamiliar appearance. This is not to claim that sentience can have no purchase whatsoever when it comes to moral ascription, but rather to assert that the vague description of sentience used in the Organic View, on my reading, provides an *anthropocentrically biased* understanding of what *constitutes* sentience in the first place. Even within biological species we still struggle to accurately discriminate between creatures that are "genuinely" capable of affect or not, often relying on anthropocentric intuitions instead of argument, as noted in the example of fish cognition above.

Epistemic issues

The second complication to be unpacked is the distinction between a mere ersatz phenomenon and its "true" instantiation. This is an idea which has a considerable amount of philosophical baggage, has been around since at least Plato, and which is a recurring theme throughout the Western philosophical canon (Gunkel 2012, p. 138). By making use of sentience as the underlying capacity which qualifies/disqualifies an entity as having a moral stake, what the Organic View is in fact claiming is that only entities with the *real* capacity for phenomenal states qualify: the mere appearance of behavioural cues that point to phenomenal states (as may be the case with anthropomorphic robots) is not enough to ground our moral ascriptions, and as such only entities that are *genuinely* sentient can be accorded a moral stake. Moreover, Torrance also claims that the type of consciousness that should ground a coherent account of moral status should track a "thick" conception of phenomenality (Torrance 2007). On this "thick" conception of phenomenality a person's consciousness is deeply embodied ("lived embodiment", to use Torrance's phrasing) and inseparable from everything about that person (ibid., p. 160). This is in contrast to "thin" conceptions of phenomenality, which Torrance takes issue with. "Thin" conceptions tend to view consciousness as something that can be detached from the entity in question. The key question then becomes how we are to go about recognizing whether entities are phenomenally conscious (sentient) in this "thick" sense.

How exactly are we to go about "proving" that an organism is sentient, *really* sentient (i.e. "phenomenally conscious")? As Dennett derisively points out, "everybody agrees that sentience requires sensitivity plus some further as yet unidentified 'factor *x*'" (1996, p. 66). Considering my discussion above regarding how we conceptualise sentience in non-human creatures, how are we to make an epistemically sound judgment as to what counts as, for example, "real" pain versus the mere "appearance" of pain? The fuzzy nature of the concept being employed (sentience) renders it immune to such an analysis.

To see how this might be the case, consider the classic British television game show *Would I Lie to You?* In the show, contestants are split into two teams, competing against one another in attempts at deception. In each round, one contestant from each team is randomly selected and reads aloud from a card with a note on it. The content of the note is unknown to the contestants until they read it, and the goal for the contestant who has read the card out loud is to convince their opponents that what has been read is in fact the truth. The content on the cards is of a personal nature, and so only the contestant who is reading the card will know whether it is the truth or not: the opponents have no idea and are allowed to ask probing questions, which the speaker must attempt to answer in a believable way. Once the questioning is over, the opposing team can decide to either claim that they believe the speaker to be telling the truth, or claim that they believe them to have lied. After they have submitted their decision, the speaker reveals whether the note was in

fact a truth or a lie, and if the opponents guessed correctly, they receive a point.

While British television can be as dry as academic philosophy, that is not the point I wish to make. In the case of the game show there is a type of deception at play: the speaker is attempting to convince the other team of the truth or falsity of their note. Similarly, when discussing questions of true sentience versus ersatz-sentience, we are attempting to figure who or what is on either side of the divide. We interpret the available evidence and then need to come to a sound judgment about the entity in question. However, and this point is crucial, in the case of the game the deception is removed: we are shown the veracity of the matter when the speaker reveals whether they were telling the truth or not. In the case of our sentience ascriptions, we have no such epistemic security: we do not have the privileged access required in order to know whether we have made the correct judgment or not. There is no verifiable test we can perform in order to determine whether we have made the correct kind of ascription. The reason we have these issues is due to a kind of epistemic opacity—we do not have direct access to the qualitative states of others and are therefore not in a good position to judge whether an entity is "truly" sentient or not.[10] While the Organic View does not explicitly require such access to these states, it nonetheless remains an open question as to how we are to know whether the entity in question is sentient. It seems as though we would also require further evidence which would, for example, show that there is indeed a *causal connection* between being a biological organism and sentience.

Moreover, as argued by Gunkel, any inference about internal states made from various external cues requires a "leap of faith" (2012). This leap, according to Gunkel, is not properly defined nor easily defendable in each and every case. An example he uses is that of a cat that screams in pain versus a lobster which is being boiled. Our intuitions suggest to us that the cat is clearly suffering, whereas things are not so clear in the case of the lobster. What Gunkel is cautioning us against here, therefore, is how we go about inducing from exterior resemblances of "suffering" (with reference to our own case) to an interior analogy regarding a seemingly coherent conception of sentence. If, as the Organic View suggests, we ought to view biological sentience as a necessary condition for moral patiency, then it should follow that we have a clear way to go about identifying the entities that do or do not meet the requirement.

Consider the advent of advanced neuroimaging technology, such as functional magnetic resonance imaging (fMRI), which allows us to detect brain activity associated with blood flow. This type of technology allows us peer into the "moving parts" in the brain which may be correlated with sentience. However, talk of internal states and the talk of how we describe, scientifically, the information that an fMRI machine represents to us are two very different language games. We therefore cannot know whether two equivalent systems—one inorganic the other organic—are phenomenally different by merely putting them through a scanner. To attempt to explain what these internal states "feel like" in terms of neurophysiology and physics would be a category mistake (Powers 2013, p. 233).

This issue precludes us from being able to use "true" sentience, as specified in the Organic View, as a qualification for moral status, whether biological or artificial. The argument I have put forward, therefore, undermines the specific notion of strictly biological sentience put forward by the Organic View. What this implies, for my purposes, is that there is now conceptual space for the notion that some future artificial system may come to have a moral stake, without necessarily being sentient in the sense specified by the Organic View. We simply don't know if various entities—including other people—are only apparently or "truly" sentient. Hence, we could decide to treat all apparently sentient creatures as moral patients, which implies at some point an AI may be worthy of this type of moral attribution.

Towards a coherent account of moral patiency

From the failure of the Organic View, I would like to tentatively suggest a model for future research into moral patiency, a model which does not operate on the same biases as the Organic View. This view is broadly in line with the social-relational account presented by Coeckelbergh (2010b, 2014). Coeckelbergh claims that both "direct" (such as those based on utilitarian or deontological criteria) and "indirect" (such as those based on virtue ethics) arguments for moral patiency rest on the ontological features of the entity in question, a feature which poses significant issues for both kinds of theory (Coeckelbergh 2010b). One aspect of Coeckelbergh's argument that I think could be further developed, however, is how we might come to determine the various mental states we deem important for moral consideration (Coeckelbergh 2010a; Torrance 2013). I believe that it is possible for a defender of social-relational approaches to

[10] Torrance does address this issue (2014) and refers to the view that I broadly defend in this paper as "social relationism" (SR). Torrance claims that SR positions do not offer us "inherently right or wrong answers" when it comes to questions of moral patiency (2014: 12). I think this a somewhat superficial reading of SR approaches, but it is beyond the scope of this paper to go into any detail in this regard, as my focus here is concerned with the specific claims made by Torrance with regards to the criteria of moral status specifically, not realism versus social relationism more generally.

draw on certain conceptual resources developed by Daniel Dennett.

Consider how we come to infer the psychological states of others on a day-to-day basis (usually without the use of advanced neuroimaging equipment): we largely use external cues in order to make plausible predications about what might be going on in their craniums. However, this type of projection is not necessarily indicative of the "real" type of phenomenal ascription required for sentience as specified above, but at the very least it provides a predictive model that we can use to infer what might be going on in other people's heads. This methodological approach, formalised by Dennett, is known as the intentional stance (1989).[11] This "intentional stance" treats the agent in question as a rational one, and then attempts to figure out which beliefs and desires the agent ought to have in light of this capacity (ibid., p. 17). Imbued in Dennett's exposition of the intentional stance is a willingness to let go of certain outdated conceptual categories. He is willing to acknowledge that mental postulates such as "beliefs", "desires", etc. are useful for predicting behaviour, but are not good guides as to what is *really going on in the brain*. They are therefore not good theoretical entities, which is why the intentional stance must remain (and is) *non-committal* (or theory neutral) with regards to the internal structures that underlie the specific competencies that an investigator is explaining (Stich 1981, p. 44; Yu and Fuller 1986, p. 454; Dennett 2009, p. 10). In this way the intentional stance is neutral on what ontological properties need to be present in the entity under investigation – so long as we can make reasonable predications regarding the entity's behaviour by *ascribing* beliefs and desires to it, we would be correct in considering it an intentional system (Slors 1996, p. 94).

Likewise, we might be able to use the intentional stance to try to determine whether an entity in question is indeed worthy of moral consideration, based on certain behavioural cues.[12] This approach should not, however, be seen as exhaustive: it is only a helpful heuristic as to whether an entity is in fact sentient. While relying *only* on behaviouristic cues would mean that we would accord a moral stake to anything capable of, for example, mimicking pain, this would be a mischaracterisation of my proposed usage of Dennett's methodology. My suggestion is simply that we take these behavioural cues seriously, and use them in conjunction with other relevant social-relational criteria, as opposed to only relying on the presumed capacity to have "real" qualitative states of experience or having a particular causal history, criteria that play key roles in the Organic View. In addition to behavioural cues, we might look to other cues indicative of an entity's internal constitution and what this tells us about the likelihood of this entity having the capacity for affect. This type of naturalistic approach is exemplified in the example of fish cognition above, in which our concepts and their associated usage are consistent with and do not contradict our best science (Ritchie 2008; Brown 2015). We might find that we over-ascribe the capacity for affect on this approach, but it is surely better to err on the side of caution when it comes to moral concern. The further value of creating a space in our moral and conceptual landscape for AAs is that by doing so we can perhaps solve so-called "responsibility-" and "retribution-gaps" (see Champagne and Tonkens 2013; Müller 2014; Gunkel 2017; Nyholm 2017). The former refers to cases in which it is unclear whether a human being or an AA was responsible for a moral action. The latter refers to cases in which AAs are involved in producing moral harms. In such scenarios people may feel a strong urge to punish somebody for the moral harm, but there may be no appropriate target for this punishment (Nyholm 2017).

Two more behaviouristic and functional approaches to moral ascription are the Moral Turing Test (MTT) (Gerdes and Øhrstrøm 2015) and Turing Triage Test (TTT) (Sparrow 2004). The first of these tests asks whether an artificial system "acts at least according to the ethical standards that are normally considered acceptable in human society" (Gerdes and Øhrstrøm 2015, p. 99).[13] If the system can pass such a test, then it can be worthy of moral consideration.[14] The TTT test proposes that in a "triage"[15] situation if one human person is replaced with an AA, and the moral character of the dilemma remains intact, then the AA would have achieved moral standing comparable to that of human beings (Sparrow 2004, p. 203). Both of the aforementioned propose novel ways in which we might come to understand the moral

[11] My decision to make use of the intentional stance is far from uncontroversial. Dennett believes that a third-person, materialistic starting point is the most appropriate one for further investigations into mentalistic concepts. This, however, can be contested on various grounds. See, for example, Nagel (1986), Ratcliffe (2001) and Slors (1996, 2015) for various philosophical issues with Dennett's account. It is far beyond the scope of the present paper to resolve these and other problems with Dennett's theory. For my purposes, however, what matters is that social-relational accounts can be amended with a theory which accounts for mental states, the details of which would still need to be worked out.

[12] These could be signs that are indicative of suffering, for example vocalizations (sighing or moaning), facial expressions (grimacing, frowning, rapid blinking, etc.) or bodily movement (being hunched over, exterior rigidity, etc.).

[13] For a critique of the Moral Turing Test, see Arnold and Scheutz (2016).

[14] Also see Wallach and Allen (2009: 70) for an exposition of the comparative Moral Turing Test (cMMT), which asks "which of these agents is less moral than the other?", as opposed to the question of which entity is the artificial agent, posed in the MTT.

[15] A situation in which a choice must be made as to which of two human lives to save.

contours of our relationships with intelligent machines in the future.

Recommendations for future research

A key issue faced by any account of moral patiency, however, is how such frameworks ought to deal with cases where the AA in question does not necessarily have humanoid features but nonetheless exhibits certain external cues that lead us to believe that it should be accorded some kind of moral concern.[16] In such cases, it is surely better to erroneously accord moral concern than to unjustifiably deny it (Wareham 2011, p. 39). A further question concerns just what exactly "machine consciousness" entails, as it need not necessarily be anything like human consciousness, making the solution to the question of machine moral patiency even more seemingly intractable. Good attempts at a philosophically coherent account of machine moral patiency are provided by Sparrow (2004), Wareham (2011), Coeckelbergh (2014), and Danaher (2017b).

In this paper I have problematised the specific sense of sentience proposed by the Organic View. To make this case I provided an exposition of the claims argued for by the Organic View, and then provided two critiques, one conceptual and the other epistemic, which served the purpose of illuminating the need for a social-relational philosophical methodology when it comes to machine moral patiency (Coeckelbergh 2010b). This new approach was introduced through the lens of Daniel Dennett's intentional stance, which could in future serve as a more philosophically coherent framework for these kinds of issues. What this implies for future research into moral patiency is that we should be careful in how we operationalise certain key concepts, such as sentience, and guard against anthropocentric fallacies as best we can. Shifting towards a social-relational methodological framework that places more emphasis on external cues might be one such way to mitigate this risk.

Acknowledgements I would like to thank my supervisor and mentor Tanya de Villiers-Botha for her insightful comments and guidance. I am also indebted to Deryck Hougaard and Lize Alberts who read earlier drafts of this paper and provided very useful feedback.

[16] Another arena requiring further research is the use and distribution of "entertainment" robots (Royakkers and van Est, 2015). More specifically, sex robots, which raise questions concerning the role of consent and ownership, and how (if it all) these concepts refer in this case. If we concede that such robots are AAs, can they give meaningful consent? Moreover, can we legitimately speak of acts such as "robotic rape", and punish those performing such acts (see Danaher, 2017a)? More work needs to be done at both the philosophical and regulatory levels to unpack solutions to these and other questions.

Reference

Arnold, T., & Scheutz, M. (2016). 'Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology, 18*(2), 103–115. https://doi.org/10.1007/s10676-016-9389-x.

Bostrom, N., & Yudkowsky, E. (2011). The ethics of artificial intelligence. In K. Frankish (Ed.), *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge University Press.

Brown, C. (2015). Fish intelligence, sentience and ethics. *Animal Cognition, 18*(1), 1–17. https://doi.org/10.1007/s10071-014-0761-0.

Chalmers, D. J. (1996). *The conscious mind*. Oxford: Oxford University Press.

Champagne, M., & Tonkens, R. (2013). Bridging the responsibility gap. *Philosophy and Technology, 28*(1), 125–137.

Coeckelbergh, M. (2010a). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y.

Coeckelbergh, M. (2010b). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221. https://doi.org/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*, 61–77. https://doi.org/10.1007/s13347-013-0133-8.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences, 15*(8), 358–364. https://doi.org/10.1016/j.tics.2011.06.008.

Danaher, J. (2017a). Robotic rape and robotic child sexual abuse: Should they be criminalised? *Criminal Law and Philosophy, 11*(1), 71–95. https://doi.org/10.1007/s11572-014-9362-x.

Danaher, J. (2017b). The rise of the robots and the crisis of moral patiency. *AI and Society*. https://doi.org/10.1007/s00146-017-0773-9.

Dennett, D. (2009). Intentional systems theory. In *The Oxford handbook of philosophy of mind*, (Dennett) (pp. 1–22). https://doi.org/10.1093/oxfordhb/9780199262618.003.0020.

Dennett, D. C. (1989). *The intentional stance*. Cambridge, Massachusetts: MIT Press. https://doi.org/10.1017/S0140525X00058611.

Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology, 1*, 37–56. https://doi.org/10.1023/A:1010018611096.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machine, 14*, 349–379. https://doi.org/10.2139/ssrn.1124296.

Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral turing test. *Journal of Information, Communication and Ethics in Society, 13*(2), 98–109. https://doi.org/10.1108/JICES-09-2014-0038.

Gunkel, D. J. (2012). *The machine question*. London: MIT Press.

Gunkel, D. J. (2017). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9428-2.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5.

Johansson, L. (2010). The functional morality of robots. *International Journal of Technoethics, 1*(4), 65–73.

Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics*. https://doi.org/10.1007/s10551-014-2180-1.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2–3), 123–133. https://doi.org/10.1007/s10676-008-9174-6.

Johnson, D. G., & Noorman, M. (2014). Artefactual agency and artefactual moral agency. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 143–158). New York: Springer.

Leopold, A. (1948) A land ethic. In *A sand county almanac with essays on conservation from Round River*. New York: Oxford University Press.

Müller, V. C. (2014). Autonomous killer robots are probably good news. *Frontiers in Artificial Intelligence and Applications, 273*, 297–305. https://doi.org/10.3233/978-1-61499-480-0-297.

Naess, A. (1973). The shallow and the deep long-range ecology movements. *Inquiry, 16*, 95–100.

Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press. https://doi.org/10.2307/2108026.

Nyholm, S. (2017). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci'. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-017-9943-x.

Powers, T. M. (2013). On the moral agency of computers. *Topoi, 32*(2), 227–236. https://doi.org/10.1007/s11245-012-9149-4.

Ratcliffe, M. (2001). A kantian stance on the intentional stance. *Biology and Philosophy, 16*(1), 29–52. https://doi.org/10.1023/A:1006710821443.

Ritchie, J. (2008). *Understanding naturalism*. Stocksfield: Acumen.

Royakkers, L., & van Est, R. (2015). A literature review on new robotics: Automation from love to war. *International Journal of Social Robotics., 7*(5), 549–570. https://doi.org/10.1007/s12369-015-0295-x.

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals*. New York: New York Review of Books.

Singer, P. (2011). *The expanding circle: Ethics, evolution and moral progress*. New Jersey: Princetown University Press.

Slors, M. (1996). Why Dennett cannot explain what it is to adopt the intentional stance. *The Philosophical Quarterly, 46*(182), 93–98.

Slors, M. (2015). Two improvements to the intentional stance theory: Hutto and Satne on naturalizing content. *Philosophia (United States), 43*(3), 579–591. https://doi.org/10.1007/s11406-015-9627-1.

Sparrow, R. (2004). The Turing triage test. *Ethics and Information Technology, 6*(4), 203–213. https://doi.org/10.1007/s10676-004-6491-2.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–78. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Stich, S. P. (1981). Dennett on intentional systems. *Functionalism and the Philosophy of Mind, 12*(1), 39–62.

Sullins, J. P. (2011). When is a robot a moral agent? *Machine Ethics, 6*(2001), 151–161. https://doi.org/10.1017/CBO9780511978036.021.

Torrance, S. (2007). Two conceptions of machine phenomenality. *Journal of Consciousness Studies, 14*(7), 154–166.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI and Society, 22*(4), 495–521. https://doi.org/10.1007/s00146-007-0091-8.

Torrance, S. (2013). Artificial agents and the expanding ethical circle. *AI and Society, 28*(4), 399–414. https://doi.org/10.1007/s00146-012-0422-2.

Torrance, S. (2014). Artificial consciousness and artificial ethics: Between realism and social relationism. *Philosophy and Technology, 27*(1), 9–29. https://doi.org/10.1007/s13347-013-0136-5.

Wallach, W., & Allen, C. (2009). *Moral machines*. New York: Oxford University Press.

Wareham, C. (2011). On the moral equality of artificial agents. *International Journal of Technoethics, 2*(1), 35–42. https://doi.org/10.4018/jte.2011010103.

Yu, P., & Fuller, G. (1986). A critique of Dennett. *Synthese, 66*(3), 453–476.