How could we know when a robot was a moral patient?

Dr Henry Shevlin, Leverhulme Centre for the Future of Intelligence

ABSTRACT. There is growing interest in machine ethics in the question of whether and under what circumstances an artificial intelligence would deserve moral consideration. In this paper, I explore a particular type of moral status that I term *psychological moral patiency*, focusing on the epistemological question of what sort of evidence might lead us to reasonably conclude that a given artificial system qualified as having this status. I survey five possible criteria we might apply, namely intuitive judgments, assessments of intelligence, the presence of desires and autonomous behaviour, evidence of sentience, and behavioural equivalence. I suggest that despite its limitations, the latter approach offers our best way forward, and defend a variant of that I term the *cognitive equivalence strategy*. In short, this holds that we should consider an artificial system to be a psychological moral patient to the extent that it possesses cognitive mechanisms shared with other beings such as non-human animals whom we also consider to be psychological moral patients.

1. Introduction

Even a casual viewer or reader of science-fiction will doubtless have encountered the image of the maltreated machine. From the abused simulants of *Blade Runner* to the neglected child-robot 'David' in *A.I. Artificial Intelligence*, we seem to have little difficulty in imagining artificial beings feeling pain and distress, and even empathising with them. The kind of advanced artificial intelligence (AI) we see in fiction is of course far removed from the capabilities of real world machines. However, as the capacities of artificial intelligence continue to improve, interest has grown within the AI ethics community concerning the question of whether – and when – artificial beings may reasonably come to possess or demand some form of moral status.[*]

This paper aims to engage with a specifically epistemological aspect of this question, namely what sort of *evidence* or *criteria* could or should be used to determine whether an artificial intelligence possesses moral status, and specifically a form of moral status I term psychological moral patiency. In the next section I briefly spell out this notion of a psychological moral patient, distinguishing it from other grounds of moral obligation. In Section 3, the main part of the paper, I consider five possible strategies we might adopt for identifying artificial psychological moral patients, namely (i) reliance on intuition and empathy, (ii) intelligence-based measures, (iii) the presence of autonomous desires, (iv) the presence of empathy, and (v) behavioural equivalence to beings to whom we already take ourselves to have moral obligations. While all of these strategies have their merits, I go on in Section 4 to argue for a variant of (v) that I term the *cognitive equivalence strategy*, which suggests that we assign moral status to artificial beings based on whether – and how closely – they instantiate cognitive capacities that our best current science attributes to other beings we already take to qualify as moral patients.

2. Moral relevance and moral patiency

In everyday ethical decision-making we typically understand ourselves to have a variety of sources of ethical obligation, including those arising from the demands of friendship, family, personal integrity, political commitments, and religious or spiritual imperatives. Among these obligations, however, one important class comprises the obligations we have to other beings in virtue of their possession of *intrinsic interests*, such as the interest in not being made to suffer or not having their autonomy violated. These obligations extend to our fellow humans

[*] For notable recent discussions of the topic, see Basl, 2014; Bryson, 2010; Coeckelbergh, 2018; Danaher, 2019; Gunkel, 2018; Neely, 2014; Schwitzgebel & Garza, 2015; Sparrow, 2004; and Tomasik, 2014.

as well as to at least some non-human animals, but not – typically at least – to inanimate objects such as rocks, cars, or houses. I will use the term moral patients to refer to beings who are capable of exerting moral obligation on us in virtue of the possession of such intrinsic interests.[*] This account of a moral patient is still quite broad, and potentially includes beings like trees, flowers, and seeds that lack psychological states like attitudes or experiences yet may possess interests relating to their biological function such as growth and development. While it is a matter of debate whether such interests alone justify the ascription of moral patiency to a being [1], I will set the matter aside in what follows, and focus specifically on what I term *psychological* moral patiency, that is, a form of moral status that may arise in virtue of a possession of specific psychological capacities such as sentience, autonomy, desires, and so on.

I would suggest that this notion of psychological moral patiency is of great relevance to much of our moral decision-making, insofar as capacities for undergoing suffering or exercising autonomy are relevant to many difficult decisions we make. The question of whether a being is a psychological moral patient is especially salient when interacting with non-human animals, as well as people in persistent vegetative states or comas whose residual capacity for psychological states may be uncertain. When David Foster Wallace famously wondered whether lobsters feel pain [2], for example, or when neuroscientists assess whether a person in a persistent vegetative state is minimally conscious [3], these inquiries can be understood as at least partly concerned with whether the being in question qualifies as a psychological moral patient.

While I take the notion of psychological moral patient to be important for ethical decision-making, I do not assume that *only* psychological moral patients exert moral obligations upon us. As noted, some living things might qualify as moral patients in virtue of their possessing biological functions that we should endeavour to respect. Likewise, one might think that some inanimate objects like rivers and mountains and sacred artefacts qualify as 'final goods' (in the sense of Korsgaard, 1983) and matter in their own right despite lacking clearly definable interests and thus failing to be moral patients in the sense given above.

It is also not obvious that we have obligations towards a being *just* in virtue of its being a psychological moral patient. One might think, for example, that it is possible to

---

[*] The questions of how to analyse notions such as moral patiency and moral status have been discussed widely. For more detailed discussion of the general notion, see, e.g., Kamm, 2007; Korsgaard, 1996; and McMahan, 2001, and for its use in relation to artificial beings see Basl, 2014; Bryson, 2018; and Floridi, 2013.

recognise a being as sentient or autonomous while denying we have actual obligations to them. To give an extreme example, a moral community might believe it has no obligations towards outsiders while nonetheless recognising them as fully capable of suffering and being otherwise harmed. Likewise, a retributive judge might determine that a heinous criminal has waived all right to moral treatment by others. More mundanely, we sometimes recognise ourselves as having additional obligations to individuals that do not simply arise as a function of their status as moral patients, for example the special obligations that we commonly take ourselves to have towards family and friends. There is thus room for slippage between the notions of being a psychological moral patient and being an entity to whom we bear moral obligations.

With these distinctions in place, I can clearly state the question to be addressed in what follows, namely what sort of evidence could lead us to reasonably believe that an artificial system should be considered to have moral status just in virtue of its possessing appropriate psychological states. This is distinct from the question of whether an artificial system could come to have moral relevance in a broader sense (as might be the case for a particularly beautiful or significant technological artefact), and at least somewhat distinguishable from the question of the specific kinds of moral obligations we might bear towards machines.

In relation to this second issue, it would be remiss not to note in passing the rich 'relation-based' account of moral obligation provided by David Gunkel (2018) and Mark Coeckelbergh (2018). This approach stresses the primacy of interactions and relations with beings in giving an adequate account of our moral obligations to them (see also Diamond, 1978, for a similar account concerning animals). In what follows, I will not pursue this relational strategy, instead loosely following what Coeckelbergh (2012) calls a "properties approach", which is concerned with how a being's possession of particular properties such as sentience or autonomy relates to their moral status. However, I leave open the possibility that a relational approach will be ultimately necessary for an adequately rich account of the nature and scope of our obligations to machines.

3. Evidence of psychological moral patiency

I now turn to the primary concern of this paper, namely what sort of evidence might reasonably lead us to believe that an artificial being qualified as a psychological moral patient, and thus as an entity capable of exerting at least one sort of moral obligation upon us.

I should note at the outset that some would affirm that *no* sort of evidence could license this conclusion, and that no artificial being could ever thus qualify as a moral patient.

This position may even seem intuitive; as Levy (2005: 393) notes, "[t]o many people the notion of robots having rights is unthinkable." It is worth distinguishing, however, between two negative positions regarding artificial psychological moral patiency. First, one might grant that any robot that was psychologically equivalent to a human would qualify as a psychological moral patient, but deny that such equivalence is in practice possible, perhaps on similar grounds to those famously raised by Searle (1980). Second, one might deny that even a robot that was psychologically identical to a human would qualify as a moral patient, perhaps because it lacked our evolutionary history or was not a living thing.

Both of these positions are worth taking seriously, although I will not engage with them in any detail in the present context.[*] However, I will note that both face challenges and involve controversial commitments. The first position, for example, will likely require (or follow from) the rejection of functionalist and computational accounts of mentality, thus making it unpalatable for many contemporary philosophers of mind, and faces famous 'slow replacement' objections in which parts of a human brain are gradually replaced by silicon components [10].[†] For its part, the second position is vulnerable to the charge of "speciesism" (in the sense of Singer, 2009).

Setting aside such doubts, let us consider some different types of evidence that might lead us reasonably conclude that a given robot or computer program constituted a psychological moral patient. Specifically, I will consider five possible routes to such a conclusion, namely intuition and empathy, intelligence, autonomy and desires, sentience, and behavioural equivalence, noting problems that arise in each case. Note that while I identify limitations of each of these routes, these are not intended as knockdown objections; what follows is intended more in the spirit of a survey than a set of dispositive arguments.

3.1 – Intuition and empathy

As noted at the outset of the present inquiry, we have little difficulty in empathising with artificial beings in fictional contexts. Consequently, one might imagine that no grievous epistemological difficulties will arise if and when we construct artificial beings with genuine

---

[*] For more detailed discussion of these positions and arguments against them, see Schwitzgebel & Garza (2015) and Danaher (2019).

[†] For a recent vigorous defence of mind-brain identity theory and criticism of functionalism, see Polger & Shapiro (2016). Note in particular their comments on AIs with psychological capacities equivalent to those of a human: "If cognitive digital computers are possible, then multiple realization is probably true and the identity theory is probably false… [however] the ambitions of artificial intelligence [do not] decide the question of multiple realization, although they surely amount to a wager on the outcome."

claim to moral status: we will simply *perceive* that they possess mental states.[*]

It certainly seems true that in many cases we immediately and automatically empathise with artificial beings. Gunkel (2018), for example, citing studies by Rosenthal-von der Pütten et al. (2013) and Suzuki et al. (2015) notes that "researchers found that human users empathized with what appeared to be robot suffering even when they had prior experience with the device and knew that it was "just a machine"." This empathising tendency is borne out by the actions and reports of those who interact with robots on a daily basis, such as soldiers and rescue workers serving alongside 'Packbots', military robots primarily used to identify and disarm improvised explosive devices. Gunkel notes that these AIs are frequently treated as 'fellow combatants', with soldiers "giving them names, awarding them battlefield promotions, risking their own lives to protect that of the robot, and even mourning their death."

However, as these examples suggest, to the extent that we rely on human empathy as our primary arbiter of whether an artificial system constitutes a psychological moral patient, we run the risk of 'false positives', misattributing moral patiency to relatively simple systems that lack any psychological basis for such a status. One might, of course, consider this in a positive light: given our disposition to over-attribute mentality to inanimate objects, it may be almost certain that we would recognise any genuine artificial psychological moral patients as such.

False positives are not free from costs, however,; as noted by Bryson (2010), inappropriate identification with robots involves "economic and human consequence of time, money and possibly other finite resources being given to a robot that would otherwise be spent directly on humans and human interaction."

Moreover, the fact that we will likely commit false positives in our empathetic identification with artificial beings does not rule out the possibility of false negatives. The specific traits and patterns of behaviour that we naturally associate with the possession of morally relevant psychological states are likely to be somewhat parochial, tracking those forms of appearance and behaviour that are readily assimilated to our own. As Schwitzgebel & Garza (2015) note, "human beings are much readier, from infancy, to attribute mental states to entities with eyes, movement patterns that look goal-directed, and contingent patterns of responsiveness than to attribute mentality to eyeless entities with inertial movement patterns and noninteractive responses. But of course such superficial features

---

[*] For similar views about our capacity to directly perceive the mental states of other humans, see Dretske (1973) and McDowell (1998).

needn't track underlying mentality very well in AI cases."

This prompts the worry that we may fail to identify genuine artificial psychological moral patients in cases where they are too alien or too exotic to elicit our standard empathetic responses. These dangers may be particularly salient in the case of AIs that lack an embodied form or easy visualization [16]. Consequently, reliance on intuition and empathy alone seems like a risky strategy for ensuring appropriate moral responses to artificial beings.

3.2 – Intelligence

A second source of information we might use to rationally ground attributions of psychological moral patiency to artificial systems is intelligence. This is clearest in the case of human-level artificial intelligence, as demonstrated by the Turing Test (1950). Were we to build a system with the complete repertoire of our cognitive abilities and with the ability to converse indistinguishably from a human being, we would have few grounds for denying it substantial moral status (however, see Sparrow, 2004).

While human-level intelligence will satisfy many as a *sufficient* condition for psychological moral patiency, however, it is far less clear that it is a necessary one. For example, most of us consider many non-human animals to be psychological moral patients capable of exerting consequent moral obligations upon us, despite falling short of our cognitive abilities in many domains. It seems conceivable and perhaps even likely, then, that the first AIs with a genuine claim to moral consideration in virtue of their psychological capacities will nonetheless lack human-level intelligence.

However, there are potentially useful intelligence-based measures of psychological moral patiency besides the Turing Test. In a future era of robotics, for example, we might administer 'intelligence tests' to artificial systems, and use these to guide our attributions of moral patiency to them. Needless to say, this is a proposal fraught with difficulties. To begin with, the history of intelligence testing as a guide to ethical decision-making is famously dubious, beset with prejudice and racial bias [19], and we should not expect the use of intelligence testing in regard to artificial systems to be immune to such forms of prejudice.

Moreover, it is worth noting the somewhat complex relationship between intelligence and psychological moral patiency. It certainly seems to be the case that we are more willing to extend moral status to animals with impressive cognitive abilities: while few would hesitate to grant moral status to dogs, dolphins, and chimpanzees, creatures with smaller and less complex nervous systems such as insects and crustaceans constitute a grey area in our folk morality. However, it is not clear that it is intelligence per se that motivates this association, so much as the inferred link between cognitive complexity and sentience (see

3.4, below). [*] And certainly, most of us would not endorse the use of relative intelligence as a primary criterion for making deciding between the ethical claims of different individual humans.

A final worry for the use of intelligence as a criterion for psychological moral patiency concerns the challenge of even giving an adequate account of the notion. While Sparrow, (2004), for example, may be right to say that "we seem to have a firm intuitive grasp of what intelligence is", translating our pretheoretical concept of intelligence into one that can be operationalised and applied to systems quite different from us is deeply challenging, as reflected by the confusion about the very meaning of the term "artificial intelligence."[†] Current artificial systems dramatically outperform humans in a variety of cognitively demanding tasks, from arithmetic to chess and (certain forms of) image recognition, yet none are plausibly as intelligent as a human or even most animals [20]. Quite apart from the normative challenges discussed above, then, any intelligence-based measure for assessing psychological moral patiency faces the unenviable if not impossible task of determining how to weight and synthesise the myriad forms of cognitive proficiency into a unified measure of a system's overall intelligence [21].

3.3 – Autonomy and desires

Rather than focus on intelligence *tout court*, we might instead attempt to use evidence of some specific aspect of a machine's intelligence as a guide to its status as a psychological moral patient. There are of course many such aspects of intelligence we might appeal to, such as its flexibility, creativity, or memory, but for present purposes, I will consider whether *autonomy* might constitute an appealing criterion. This is a suggestion persuasively developed by Neely (2014), who identifies autonomy as a measure for whether a system possesses genuine desires which in turn would ground moral status.

> [C]onsider the case where the agent's goals are not always determined by an outside source, i.e., where the agent is capable of determining its own goals at least some of the time. In this case, the agent is expressing a basic capacity for autonomy, which implies that these goals must be chosen by the agent itself.

---

[*] Note, for example, that the British Animals (Scientific Procedures) Act of 1986 which previously protected all vertebrate animals was extended by amendment in 1993 to include octopuses. This amendment was made not on the grounds of intelligence per se, but rather because the complex nervous system of the octopus created reasonable grounds for inferring that octopuses might feel pain.

[†] See Wang (2019) for a thorough discussion of this topic.

The possession of such desires, she argues, gives us grounds for assigning a system some kind of moral status; "while the agent's desires may be overridden, they may not simply be ignored."

There are interesting similarities between Neely's position and some leading positions in animal welfare ethics. Dawkins (2017), for example, argues that we can construct a science of animal welfare based on the satisfaction of preferences. If an animal has very strong preferences that are not being met – for food, mating, or a certain kind of environment, say – then their individual interests are not being satisfied. As she puts it, "'having what they want' is a shorthand way of covering a wide variety of way in which animals can, in a publicly observable way, show us whether the environment they are in is positive (something they like and want to continue with) or negative (something they dislike and want to escape from or avoid)" (Dawkins, 2017).

The positions of both Neely and Dawkins are well-developed and sophisticated, and a detailed response lies beyond the scope of the present paper. Moreover, Neely acknowledges the difficulties involved in assessing whether a given system is truly autonomous or not, suggesting that we should treat such as assessments as probabilistic. Nonetheless, I would note that the use of autonomy as a criterion for moral patiency faces some key worries. For one, as Neely herself notes, it is not a necessary condition for patiency: a system may be sentient and thus capable of suffering despite lacking endogenous desires (perhaps having its attitudes directly 'programmed in').

Perhaps a more fundamental issue, however, comes from the fact that there are many beings that at least superficially exhibit autonomy yet lack any claim to serious moral consideration. The aversive responses of nematode worms, the chemotaxis of bacteria, and the phototaxis of plants, for example, could all be loosely classified as autonomous desires in the broadest sense. Even if we take these behaviours as constituting some minimal claim for moral status, it is a claim that would be outweighed by even fairly trivial instrumental considerations. To use an example from Basl (2014), were there to be some scientific interest to be gained from growing and destroying a million maple trees, few would balk at the experiment on moral grounds, assuming it generated no significant negative externalities.

What is required for an account of moral patiency founded in autonomous desire to have real ethical force is some means of distinguishing between mere appetitive and aversive behaviours and robust preferences. Some philosophers in the preference utilitarian tradition (including Singer, 1979, and Varner, 2012) limit moral patiency to conscious or sentient beings with preferences, but this would of course be at odds with Neely and Dawkins' goal of

providing a separate criterion for moral patiency.[*]

It is of course possible that ongoing work in cognitive science may identify a principled criterion for distinguishing morally salient 'thick' preferences from simpler forms of appetitive and aversive behaviour and learning. However, absent such a criterion, an approach to moral patiency founded on the use of autonomy as a guide to possession of morally significant desires is at best incomplete.

3.4 – Sentience

Perhaps the most obvious criterion for artificial psychological moral patiency is sentience, defined in the present context as a capacity for undergoing conscious states like pleasure and pain that feel good or bad to them; in the terminology of contemporary psychology, a capacity for undergoing positively and negatively valenced conscious states. Even if we do not wish to claim that this is 'all there is' to psychological moral patiency, certainly it is natural to think that if a being can feel pain, it has at least some potentially morally salient interests, namely *not* feeling pain. Applying this to artificial entities, then, the notion of sentience might provide a conceptually straightforward pathway to moral patiency: if and when we build machines capable of consciously experiencing pleasure, pain, or other valenced states, we will have grounds for recognising them as moral patients.

There are several attractive features of this approach. First, as noted, the claim that sentience grounds at least some minimal form of moral patiency is relatively uncontroversial; even if we wish to allow there is more to well-being than pain and distress, it is hard to deny that such experiences are often detrimental to our interests and those of non-human animals. Second, a sentience-based approach to establishing moral patiency can accommodate quite simple forms of sentient being. Thus while we may not wish to consider fish or lobsters *persons* or as exhibiting autonomy in a rich sense, if it could be established that they could feel pain this would give us some grounds for potentially taking their interests into account. And as noted above, it seems antecedently possible or likely that long before we have achieved human-level artificial intelligence with claims to more robust moral notions such as personhood or self-determination, we will along the way build systems with the cognitive complexity equivalent to that of non-human animals that might potentially qualify as moral patients in virtue of capacities to experience pain, fear, or other simple forms of negatively

---

[*] It should be noted that Neely herself would likely endorse the adoption of extremely broad notions of autonomy and desire, acknowledging that her proposals involves "a large expansion to the moral community". As suggested by Basl's example of the maple trees, however, this faces the worry that any moral obligations consequent upon such expansive notions would have little practical moral relevance.

valenced state. Finally, a sentience-based approach to psychological moral patiency seems at least in principle empirically tractable, insofar as it identifies this status with the capacity to undergo specific kinds of psychological state which we might expect to be amenable to assessment via the tools of cognitive science.

Despite this superficial simplicity, however, sentience-based approaches to moral patiency face serious challenges of both a philosophical and practical character. The most daunting is of course the challenge of how we can ever establish whether a given system is conscious. This is an area of intense debate within comparative psychology, but expert opinions vary wildly; some philosophers and scientist extend consciousness to all vertebrates and some invertebrates [26], while others have suggested that organisms as relatively cognitive sophisticated as fish may lack the capacity to feel conscious pain [27]. There are even those (including Carruthers, 2018) who believe the very question of non-human consciousness to be ill-posed.

However great the challenges are for assigning consciousness to animals, the corresponding challenges for assigning it to machines will be far greater. In the case of animals, we can at least appeal to broad similarities in cognitive architecture and material composition, as well as a shared evolutionary history. By contrast, most artificial systems possess – and will continue to possess – wildly divergent forms of cognitive architecture and processing substrate from those found in human beings.

Such controversies lead some including Gunkel (2018), to a pessimistic conclusion about the prospects of a science of artificial consciousness. As he puts it, the present science of consciousness is "unable to demonstrate with any certitude whether animals, machines, or other entities are in fact conscious (or sentient) and therefore legitimate moral persons (or not), [and] we are left doubting whether we can even say the same for other human beings."

Such extreme pessimism may not be warranted; considerable progress has been made in the last few decades on the development of better scientific theories of consciousness, with practical implications including better tools for the assessment of consciousness in patients in vegetative states [29]. Nonetheless, in light of the fundamental theoretical disputes ongoing among the many quite different scientific approaches to consciousness [30], even the optimist must grant that a settled consensus in the field remains a distant goal. Consequently, and in light of the rapid progress in the development of artificial intelligence, we should not too much hope in the idea that we will be able to rely on scientific measures of sentience to establish psychological moral patiency in artificial systems.

3.5 – Behavioural equivalence

Faced with the daunting challenges of assessing the occurrence of states like consciousness in other beings, a final recourse may be to appeal to some notion of *behavioural equivalency*. This is the approach developed by Danaher (2019) in the context of what he calls a theory of *ethical behaviourism*. Danaher claims that in our ordinary ascriptions of moral status to other beings (including other humans), we rely fundamentally on behaviour as a source of evidence: even if someone takes moral status to involve capacities such as sentience or a soul, their evidence for ascribing these capacities, he argues, will ultimately derive from their observations of behaviour. On this basis, Danaher argues for a criterion of *performative equivalence* in the assessment of artificial moral patiency: "if a robot consistently behaves as if it is in pain, and if the capacity to feel pain is a ground of moral status, then a robot should be granted the same moral status as any other entity to whom we ascribe moral status on the grounds that they can feel pain."

Behavioural equivalence as an approach to artificial psychological moral patiency has attractive features. It is not hostage to the fortunes of any future advances in cognitive science, but is instead immediately applicable: for any candidate artificial moral patient, we can simply ask whether its behavioural repertoire is equivalent to that of an entity to which we already grant the status of moral patiency. Likewise, it avoids some of the pitfalls of a purely intuitive approach, insofar as it relies not on our immediate empathetic responses to artificial beings, but on the (in principle) more rigorous criteria of behavioural similarity.

Danaher's position is again complex and well-developed, and the view I will ultimately defend below draws heavily upon it. Nonetheless, I would suggest that it has features that make it at least somewhat problematic. For one, it is limited in application to systems whose behaviour is at least somewhat similar to that of beings to which we already attribute moral status: when no behavioural equivalency can be found, it must at best remain agnostic about questions of psychological moral patiency. For another, it is at risk of being 'gamed': given some criterion for behavioural equivalence (say, avoidance behaviour), a dedicated computer scientist may be able to produce a robot or system with the capability of producing the relevant behaviour yet doing nothing else.[*] In such a case, we might have little reason to consider the system to have any moral status, yet it would have satisfied our earlier criteria.

These problems relate to a deeper one that Danaher himself notes, namely the

---

[*] This is a broader worry for theories of moral patiency that operationalise the morally relevant psychological capacity in terms of some specific behaviour of information-processing capability. As Tomasik (2014) notes, "[w]hen we develop a simple metric for measuring something… we can game the system by constructing degenerate examples of systems exhibiting that property."

challenge of determining the appropriate "performative threshold" for behavioural equivalency, in other words, how to determine the appropriate degree of generality or specificity to adopt when seeking to establish equivalence. For example, if a given stimulus applied to a robot served the function of negatively reinforcing a given behaviour, but did not produce any outward signs of distress, should we consider it equivalent to a 'punishment' or other negative state undergone by a human or animal? As Danaher notes, an excessively liberal threshold risks including systems within our moral circle inappropriately, while an excessively conservative one carries the danger that we will fail to identify morally significant harms.

This challenge for the theory illustrates perhaps its key limiting factor, namely its commitment to reliance on behaviour alone. Danaher stresses that he intends the term to be understood broadly, and to include "all external observable patterns, including functional operations of the brain." However, what it presumably cannot include is the theoretical vocabulary of cognitive science; things such as episodic memory, metacognitive representation, and affective states are not themselves entities that we can directly observe, but are instead mechanisms posited within the context of specific psychological theories. Yet I would suggest that it is by precisely by reference to these mechanisms that we can begin to adjudicate questions about the appropriate performative threshold to adopt in assessing behavioural equivalence.

To give a toy example from comparative psychology, imagine that we find a deep sea worm whose 'pain behaviour' is at least superficially similar to that of a creature we already consider to be a psychological moral patient – let us say a fish. The worm, we can suppose, recoils from damaging stimuli, and learns to avoid them in future. Is this enough to warrant ascribing it moral status? The answer may not be readily settled with appeal to behaviour alone, but instead depend on our best theories of the creature's internal cognitive architecture. If we were to determine, for example, that the behaviour was accomplished entirely by peripheral mechanisms in the skin of the worm that did not (again, according to our best theory of the cognitive structure of the organism) communicate with its central nervous system, this might push us towards a negative answer. If, by contrast, the relevant stimulation was encoded in central short- and long-term memory stores, we might take more seriously the possibility that the creature underwent some morally significant negative event. [*]

---

[*] It is possible that Danaher would agree on the importance of such forms of theoretical explanation, and would merely wish to reconstruct them in terms of expected outcomes on behaviour. If so, there is perhaps little disagreement between us, although given that such reconstructions are at least in principle possible for most of

Addressing such interpretative challenges is likely to be particularly important if we wish to determine the occurrence of potentially morally significant events in artificial systems that differ from us considerably in respect of their behavioural repertoire. To give another toy example, imagine that an intelligent artificial system responded to externally damaging stimuli by dumping the contents of its short term memory. Could this constitute a form of distress or some other morally significant mental event? To address this issue, we would again surely need to ask questions about the structure and dynamics of the system's cognitive architecture. For example, does the system have a centralised process for registering external events and responding to them? Does it possess a system of representations allowing it to 'weigh up' the predicted occurrence of such events with other possible positive and negative inputs? While such questions would of course not settle the issue of whether the specific event had any moral significance, in the context of a broader theory of the cognitive function of the system they could certainly serve to inform our considered judgment of the matter.

4. The Cognitive Equivalence Strategy

The five epistemic strategies described above all constitute possible sources of evidence that a given artificial system might qualify as a psychological moral patient. Though I have noted problems with each, I will reiterate that I do not take anything I have said to be decisive: all could be useful components of a broader toolkit for assessing the moral status of future artificial systems.

However, of the five strategies considered, I would suggest that Danaher's notion of behavioural equivalence provides the most directly promising path forward. In this final part of the paper, I wish to suggest how we might amend Danaher's proposal so as to answer some of the worries just raised. Specifically, I wish to argue for a heuristic that I will term the *cognitive equivalence strategy*. In short, this states that we should we treat an artificial system as morally significant to the extent that our best science of its cognitive structure and dynamics attributes to it psychological capacities present in other beings to which we already assign moral status.

This proposal clearly draws heavily upon Danaher's view, but as I describe below, I believe it has some advantages. Before that, however, I would make some general observations about the cognitive equivalence strategy. The first is that it should be seen very much as a moral heuristic or rule-of-thumb: it is not a *theory* of psychological moral patiency,

---

the theoretical vocabulary of cognitive science (for example via Ramsey sentences, as suggested by Carnap, 1950), it suggests the term ethical *behaviourism* for his view is at least misleading.

but a *procedure* we can apply that will allow us to make fallible but justified positive identifications of the moral status of an artificial being. Second, the theory is deliberately ecumenical and (relatively) theoretically-neutral. It does not single out any one cognitive capacity – such as sentience, autonomy, self-awareness, and so on – as uniquely relevant for moral status, but considers them all as potential sources of evidence for psychological moral patiency. It should thus hopefully be acceptable to theorists with a broad range of positions. Finally, I would suggest that it is a strategy that should be seen as tightly coupled to contemporary cognitive science, and it is intended to accommodate new developments in our understanding of potentially morally significant psychological capacities in different biological or artificial beings.

To spell out the strategy in a little more detail, then, imagine that we are considering whether some intelligent robot qualifies as an artificial psychological moral patient. The first question we would ask in this regard is what cognitive capacities we can attribute to the system. Does it possess working memory, a theory of mind, or metacognition? Can it engage in 'mental time travel' and creative problem-solving? Does it possess an internal representational system for registering 'desirable' and 'undesirable' events?

Answering these questions will be far from trivial, particularly if the system has not been fully programmed per se, but instead developed 'organically' using machine learning techniques that render its internal structure somewhat opaque [32]. Nonetheless, just as comparative psychologists can make defeasible but informed judgments about the cognitive capacities of non-human animals, so too can we hope to make similar estimates about the capabilities of the artificial system. With such estimates in hand, we can now compare the system's capacities to those of biological organisms, asking how its cognitive abilities compare to those of a rat, pigeon, or insect.[*] If and to the extent that the artificial system possesses relevantly similar cognitive capacities to those of such beings, we can then tentatively assign it moral status comparable to that we assign to them.

In practice, of course, such comparisons will be awkward and incomplete. It is very unlikely that any future artificial system will possess the exact same cognitive mechanisms of, say, a dog, unless deliberately constructed to do so. Instead, we might expect it to exhibit greater cognitive abilities in some domains and less in others. This prompts the awkward question of which capacities matter and to what extent. While I will not attempt to adjudicate

---

[*] Comparisons between the capabilities artificial beings and non-human animals is an exciting and active area of research. Consider, for example, the recent Animal-AI Olympics (http://www.animalaiolympics.com/), that pitted AIs against a set of canonical tasks from animal cognition.

this issue in any detail here, I would suggest that in practice such difficulties may be overcome via careful experiment and theory. Imagine, for example, that we believe chickens to be psychological moral patients. We can then ask which of their cognitive capacities, if absent in a specific chicken, would lead us to reasonably reduce our credence in their possession of this status. If a brain-damaged chicken were not to have mating instincts, for example, or lacked normal drives for hunger and thirst but was otherwise behaviourally identical to its conspecifics, we would not intuitively consider its moral status to be significantly different from that of other chickens. If, by contrast, it did not possess capacities for social reasoning, goal-directed agency, or had a dramatically diminished capacity for learning, we might revise downward our credences concerning its status as psychological moral patient.

We could also use the study of the range of behaviours available to decapitated or decerebrated animals (who are unlikely to be conscious or otherwise qualify as moral patients) to constrain the relevant cognitive parameters of comparison. For example, it was demonstrated by Ferrier (1886) that a decapitated frog will still rub its thigh in response to the application of a drop of acetic acid, while Gentle (1997) showed that decerebrate chickens would still modulate pecking behaviour in response to an injured beak if the injury preceded the removal of their forebrain, suggesting such behaviours are accomplished by relatively low-level mechanisms. Via a combination of empirical and reflective methods such as these, we might reasonably attempt to establish a set of cognitive dimensions relevant for moral patiency.

I recognise that the cognitive equivalence approach introduces complications not present in Danaher's behavioural equivalence account, and in particular, relies to a greater extent on the explanatory success of cognitive and comparative science. However, I would suggest that by licensing us to use broader forms of scientific explanation, it makes some otherwise opaque questions tractable. For example, as suggested by the arguments given in the preceding section, it allows us to move beyond behaviour as a guide to moral equivalence, and instead to look for fine-grained comparisons at the level of cognitive mechanisms.

This in turn may make it easier (though still far from trivial) to settle issues concerning appropriate equivalence standards. Thus we may be able to appeal to the best current scientific theories and tests of things like memory, desire, agency, and motivation to determine if a system *really* has the relevant capacities, or simply exhibits behaviour superficially indicative of possession of these. For example, episodic memory is defined by some psychologists as the ability to generate unified representations encoding the identity of

an object together with the spatial location and time at which it was encountered, also known as the 'what, where, when' criterion [35]. Such definitions (and associated experimental measures) can be utilised by the cognitive equivalence strategy, thus giving us at least an initial way of determining the appropriate degree of specificity to adopt for the purposes of determining moral commensurability.

Additionally, the cognitive equivalence strategy may allow us an entry point into assessing moral patiency in quite exotic forms of artificial intelligence whose behaviour is radically different from that of humans and animals. As noted in the example of the 'memory dumping' robot above, these cases may be hard or impossible to adjudicate on the basis of behaviour alone, but by extending our theoretical vocabulary to include posited psychological mechanisms, we may be able to identify the internal processes of exotic forms of intelligence as falling under a broader cognitive category such as the registration of positive and negative valence that is realised in beings we regard as moral patient, enabling fruitful comparisons of their respective moral standing.

I recognise, of course, that the cognitive equivalence strategy has important limitations. In particular, it relies on our existing judgements about whether various non-human animals constitute psychological moral patients. There is no reason to think that these assessments are infallible, or even particularly reliable, as suggested by the varied norms and legal protections enjoyed by different species in different jurisdictions. However, optimistically, one might hope that as comparative psychology progresses and greater attention is paid to animal welfare issues, these assessments will improve and become reliable. Via the cognitive equivalence strategy, such progress could be immediately and directly applied to our consideration of the moral status of machines.

A final question worth noting is what the cognitive equivalence strategy would say about *existing* AIs. After all, it is not uncommon to read in machine learning journals of systems that exhibit capacities like agency, creativity, and even theory of mind. Should such ascriptions lead us to assign moral status to systems already in existence, as suggested by Tomasik (2014)? Alas, a detailed response to this worry lies beyond the scope of the present paper, but I have argued elsewhere that such ascriptions should not necessarily be taken at face value [36], and that in practice, the cognitive capabilities of current artificial systems pale in comparison to those of even simpler animals [20]. Nonetheless, it does not seem outlandish to speculate that the current gulf between artificial and non-human biological intelligence may recede quickly in light of new developments, and that the cognitive equivalence strategy might soon give us reason to give defeasible and perhaps probabilistic weight [37] to the claim that some artificial beings qualify as psychological moral patients.

5. Conclusion

In this paper, I have sought to address an epistemological question that I take to be of great relevance to our ethical treatment of future artificial systems, namely what sorts of evidence and which criteria we should use to determine whether they qualify as psychological moral patients. I suggested that none of the methods considered – including reliance on intuition, on intelligence, and on demonstrations of autonomy and sentience – offers an uncomplicated answer to the question. I argued that Danaher's behavioural equivalence strategy offered a promising path forward, but offered a somewhat different formulation of the view in terms of cognitive equivalence that I believe overcomes some of its limitations.

REFERENCES

1. Basl J. Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philos Technol*. 2014. doi:10.1007/s13347-013-0122-y

2. Wallace DF. *Consider the Lobster: And Other Essays*. New York: Little, Brown; 2005.

3. Demertzi A, Schnakers C, Ledoux D, et al. Different beliefs about pain perception in the vegetative and minimally conscious states: a European survey of medical and paramedical professionals. *Prog Brain Res*. 2009;177(C):329-338. doi:10.1016/S0079-6123(09)17722-1

4. Korsgaard CM. Two Distinctions in Goodness. *Philos Rev*. 1983;92(2):169. doi:10.2307/2184924

5. Gunkel DJ. The other question: can and should robots have rights? *Ethics Inf Technol*. 2018. doi:10.1007/s10676-017-9442-4

6. Coeckelbergh M. Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering. *Kairos J Philos Sci*. 2018. doi:10.2478/kjps-2018-0007

7. Diamond C. Eating Meat and Eating People. *Philosophy*. 1978. doi:10.1017/S0031819100026334

8. Coeckelbergh M. *Growing Moral Relations: Critique of Moral Status Ascription*.; 2012. doi:10.1057/9781137025968

9. Searle JR. Minds, brains, and programs. *Behav Brain Sci*. 1980. doi:10.1017/S0140525X00005756

10. Chalmers DJ. The singularity: A philosophical analysis. *J Conscious Stud*. 2010. doi:10.1002/9781118922590.ch16

11. Singer P. Speciesism and moral status. *Metaphilosophy*. 2009. doi:10.1111/j.1467-9973.2009.01608.x

12. Rosenthal-von der Pütten AM, Krämer NC, Hoffmann L, Sobieraj S, Eimler SC. An Experimental Study on Emotional Reactions Towards a Robot. *Int J Soc Robot*. 2013. doi:10.1007/s12369-012-0173-8

13. Suzuki Y, Galli L, Ikeda A, Itakura S, Kitazaki M. Measuring empathy for human and robot hand pain using electroencephalography. *Sci Rep*. 2015. doi:10.1038/srep15924

14. Bryson JJ. Robots should be slaves. In: ; 2010. doi:10.1075/nlp.8.11bry

15. Schwitzgebel E, Garza M. A Defense of the Rights of Artificial Intelligences. *Midwest Stud Philos*. 2015. doi:10.1111/misp.12032

16. Tomasik B. *Do Artificial Reinforcement-Learning Agents Matter Morally?*; 2014.

17. Turing A. Computing Machine & Intelligence. *Mind*. 1950;LIX(236):433-460. doi:10.1093/mind/LIX.236.433

18. Sparrow R. The turing triage test. *Ethics Inf Technol*. 2004. doi:10.1007/s10676-004-6491-2

19. Gould SJ. *The Mismeasure of Man*. First edition. New York : Norton, [1981] ©1981; 1981. https://search.library.wisc.edu/catalog/999469689202121.

20. Shevlin H, Vold K, Crosby M, Halina M. The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. *EMBO Rep*. September 2019:e49177. doi:10.15252/embr.201949177

21. Hernández-Orallo J. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press; 2017.

22. Neely EL. Machines and the moral community. *Philos Technol*. 2014. doi:10.1007/s13347-013-0114-y

23. Dawkins MS. Animal welfare with and without consciousness. *J Zool*. 2017;301(1):1-10. doi:10.1111/jzo.12434

24. Singer P. Practical ethics - singer.pdf. 1979.

25. Varner GE. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two Level Utilitarianism*. Oup Usa; 2012. doi:10.1093/acprof:oso/9780199758784.001.0001

26. Barron AB, Klein C. What insects can tell us about the origins of consciousness. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol 113. ; 2016:4900-4908. doi:10.1073/pnas.1520084113

27. Key B. Fish do not feel pain and its implications for understanding phenomenal consciousness. *Biol Philos*. 2015;30(2):149-165. doi:10.1007/s10539-014-9469-4

28. Carruthers P. Comparative psychology without consciousness. *Conscious Cogn*. 2018;63:47-60. doi:10.1016/j.concog.2018.06.012

29. Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD. Detecting Awareness in the Vegetative State : Supporting Information. *Science (80-)*. 2006;313(5792):1402-1402. doi:10.1126/science.1130197

30. Block N. Comparing the major theories of consciousness. In: *The Cognitive Neurosciences*. ; 2009:1111-1122.

31. Danaher J. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Sci Eng Ethics*. 2019. doi:10.1007/s11948-019-00119-x

32. Selbst AD, Barocas S. *The Intuitive Appeal of Explainable Machines*. Vol 87. Rochester, NY; 2018. https://papers.ssrn.com/abstract=3126971. Accessed November 12, 2018.

33. Ferrier D. *The Functions of the Brain*. New York: G.P. Putnam's Sons,; 1886.

34. Gentle MJ. Pain-related behaviour following sodium urate arthritis is expressed in decerebrate chickens. *Physiol Behav*. 1997;62(3):581-584. doi:10.1016/S0031-9384(97)00164-9

35. Nairne JS. The three "Ws" of episodic memory: What, when, and where. *Am J Psychol*. 2015;128(2):267-279. doi:10.5406/amerjpsyc.128.2.0267

36. Shevlin H, Halina M. Apply rich psychological terms in AI with care. *Nat Mach Intell*. 2019;1(4):165-167. doi:10.1038/s42256-019-0039-y

37. Agar N. How to Treat Machines that Might Have Minds. *Philos Technol*. 2019. doi:10.1007/s13347-019-00357-8

38. Korsgaard CM. *The Sources of Normativity*. Vol 110. Cambridge: Cambridge University Press; 1996. doi:10.1017/cbo9780511554476

39. Kamm FM. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford, New York: Oxford University Press; 2007. doi:10.1093/acprof:oso/9780195189698.001.0001

40. McMahan J. *The Ethics of Killing: Problems at the Margins of Life*.; 2001. https://philpapers-org.ezp.lib.cam.ac.uk/rec/MCMTEO-10. Accessed May 31, 2019.

41. Bryson JJ. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics Inf Technol*. 2018. doi:10.1007/s10676-018-9448-6

42. Floridi L. *The Ethics of Information*. Oxford: Oxford University Press; 2013.

43. Polger TW, Shapiro LA. *The Multiple Realization Book*. Oxford University Press; 2016.

44. Dretske FI. Perception and Other Minds. *Noûs*. 1973. doi:10.2307/2216182

45. McDowell JH. *Meaning, Knowledge, and Reality*. Harvard University Press; 1998.

46. Wang P. On Defining Artificial Intelligence. *J Artif Gen Intell*. 2019;10(2):2019-2022. doi:10.2478/jagi-2019-0002

47. Carnap R. Empiricism, semantics, and ontology. *Rev Int Philos*. 1950;4(11):20-40.