What Matters for Moral Status: Behavioural or Cognitive Equivalence?

John Danaher, NUI Galway

Forthcoming in the *Cambridge Quarterly Review of Healthcare Ethics*,

doi:10.1017/S0963180120001024

Abstract

Henry Shevlin's paper—"How could we know when a robot was a moral patient?" – argues that we should recognize robots and artificial intelligence (AI) as psychological moral patients if they are cognitively equivalent to other beings that we already recognize as psychological moral patients (i.e., humans and, at least some, animals). In defending this cognitive equivalence strategy, Shevlin draws inspiration from the "behavioral equivalence" strategy that I have defended in previous work but argues that it is flawed in crucial respects. Unfortunately—and I guess this is hardly surprising—I cannot bring myself to agree that the cognitive equivalence strategy is the superior one. In this article, I try to explain why in three steps. First, I clarify the nature of the question that I take both myself and Shevlin to be answering. Second, I clear up some potential confusions about the behavioral equivalence strategy, addressing some other recent criticisms of it. Third, I will explain why I still favor the behavioral equivalence strategy over the cognitive equivalence one.

1. Introduction

Henry Shevlin's paper — "How could we know when a robot was a moral patient?" — provides a thoughtful analysis of a fascinating question.[i] Human moral concern is often quite parochial: traditionally, we only extended moral respect and courtesy to our closest peers (family, tribe, nation, gender etc). In more recent times, a series of moral revolutions has expanded the circle of moral concern beyond its parochial boundaries.[ii] Most of us now accept that all humans deserve equal moral respect; some of us take the idea of animal rights seriously. Will the revolution continue? Will we extend moral respect to artificial beings? What evidence would we call upon to answer that question? This last question, in particular, is what Shevlin sets out to answer.

In doing so, Shevlin defends something he calls the “cognitive equivalence” strategy. He says that we should recognise robots and artificial intelligences as *psychological moral patients* if they are cognitively equivalent to other beings that we already recognise as psychological moral patients (i.e. humans and, at least some, animals). In defending this cognitive equivalence strategy, Shevlin draws inspiration from the “behavioural equivalence” strategy that I have defended in previous work.[iii] Nevertheless, although he sees some merit in my approach, he criticises it for a number of shortcomings and defends the superiority of the cognitive equivalence strategy.

I am grateful to Shevlin for the care and attention he pays my previous work. Unfortunately — and I guess this is hardly surprising — I cannot bring myself to agree that the cognitive equivalence strategy is the superior one. In the remainder of this article, I will try to explain why. I will do so in three stages. First, I will try to clarify the nature of the question that I take both myself and Shevlin to be answering. Second, I will clear up some potential confusions about the behavioural equivalence strategy, addressing some other recent criticisms of it. Third, I will explain why I still favour the behavioural equivalence strategy over the cognitive equivalence one. I will suggest that there may not be much disagreement between myself and Shevlin in practice but that, to the extent that there is, that disagreement should be resolved in favour of the behavioural equivalence strategy.

2. What’s the problem?

Both Shevlin and I are concerned with a common problem. We are both interested in the conditions under which artificial beings might attain some kind of moral status such that they can no longer be treated as mere things and must instead be owed some basic moral respect and moral concern. More precisely, we are interested in the *evidence* that we might call upon to determine whether an artificial being is, in fact, a psychological moral patient.

This may sound straightforward enough but it is important to distinguish myself and Shevlin’s project from other, related, projects. In any debate about moral status, there are three main issues with which to contend:

A - *The ontological grounding of moral status* - What properties or attributes must a being possess in order to count as a moral patient? Is it sentience? Robust preferences and desires? Some sense of self or personhood? The capacity for moral agency? There are many candidate theories out there. Most, as Shevlin implicitly notes, focus on psychological or mentalistic properties/attributes.

B - The evidence of ontological grounding - How could we know whether a being possessed the relevant properties or attributes? What kinds of evidence point to the existence of sentience, robust preferences, personhood, agency and so on? Can we agree on some common set of markers that provide warrant for these properties?

C - The inference to ontology - How do we go from the evidence to the presumption that the ontological properties that ground moral status are present? Do we follow some formal procedure or test? Is there a risk that we might overinterpret the available evidence? Is the inference warranted by the available evidence?

All three issues are important. Lots of philosophical ink has been spilled in relation to issue A. But the debate about A is often conflated with the debate about B and C. Both myself and Shevlin are concerned with B and, to some extent, C. We care about the kinds of evidence that are relevant to deciding moral status and the process one must follow in order to infer moral status from the available evidence. Shevlin says that the relevant evidence is evidence of cognitive equivalence between a being whose moral status recognised and one whose moral status is undetermined; I claim that the relevant evidence is evidence of behavioural equivalence.

Even though we are focused on issue B, neither myself nor Shevlin can be completely neutral with respect to A. In order to make defensible claims about the kinds of evidence that would warrant an inference to moral patiency you need to have some view on the ontological grounding of moral patiency. Without that, you won't know where to look. Both myself and Shevlin are cognisant of this in our work. We both accept that the most widely accepted theories of moral patiency assume a psychological or mental grounding. We both accept that there is, however, some disagreement about the precise psychological or mental property (or set of properties) that is most significant. Neither myself nor Shevlin takes a firm view on this. I hesitate to speak for him but, for my part, this is because I do not think it matters too much which ontological theory is taken to be correct. As long as it is mentalistic or psychological in nature, the behavioural equivalence strategy will work.

3. Clarifying the Behavioural Equivalence Approach

In my previous work, I haven't used the term 'behavioural equivalence'. I have, instead, argued for a theoretical position I call *ethical behaviourism*.[iv] According to ethical behaviourism, a sufficient ground for believing that an entity has moral status is that entity being roughly behaviourally equivalent to another entity of whose moral status we are already convinced. To put it glibly: if it acts like (another) moral patient then you are (probably) warranted in believing that it is a moral patient. Behaviour is interpreted broadly under this theory — as Shevlin notes in his discussion — to include all external signs and patterns that an entity might display. This could include functional patterns of activity in the body or brain. That said, as I will argue below, there are reasons for thinking that this kind of evidence is not particularly reliable when it comes to making inferences to moral status.

Ethical behaviourism is a theoretical stance about the kinds of evidence we can use to make inferences to moral status. It is not, to repeat the point made above, a theory about the ontological grounding for moral status. Recent critics of ethical behaviourism, including Sven Nyholm[v] and Jilles Smids,[vi] have challenged ethical behaviourism on this point. In doing so I believe that they conflate the distinction between issues A and B that I outlined above. Nyholm, for example, argues that even if behavioural evidence is relevant in determining whether an entity has moral standing it does not follow that ontological properties are irrelevant to moral standing. To be more concrete, he takes issue with my claim that if an entity consistently and convincingly acts as if it has sentience (or whatever)[vii] then you are warranted in believing that it is sentient and you don't need to worry about whether it really is sentient, on the inside. He argues that what that entity can feel — on the inside — is what ultimately matters. Behaviour is not the sine qua non of moral standing. But the ethical behaviourist can agree with this view. They can agree that sentience is the ultimate ontological grounding for moral status. They just qualify this by arguing that behavioural evidence is the best way to make inferences to the presence of sentience. Ethical behaviourism is, in other words, methodologically behaviourist and not ontologically behaviourist.

In practice, one applies ethical behaviourism by using comparisons and analogies. You take an entity of whose moral status you are already convinced and compare it to another an entity whose moral status is contested and uncertain. Ideally, you repeat this for multiple comparisons in order to reach a reasonably robust conclusion about the moral status of the uncertain entity. In some cases, extensive comparison may not be possible and the moral status of the entity may remain uncertain or be subject to a highly tentative/defeasible assessment. Although ethical behaviourism is practically applied through comparison and analogy, this does not mean that the case for ethical behaviourism depends on an argument from analogy. On the contrary, and to repeat myself, ethical behaviourism is a general theory about the kinds of evidence that are relevant to determining moral standing. From this, you can derive an argument from analogy that makes the case for a particular being's moral standing — e.g. "robot X acts convincingly and consistently like a human, in all important respects, so it probably has moral status" — but you can also derive other kinds of argument, such as an inference to best explanation — "robot X consistently displays behaviours Y and Z and the best explanation for this is that it has the property of sentience which establishes moral status". [viii]

Related to this, Smids challenges my defence of ethical behaviourism for claiming to be neutral or agnostic with respect to the ontological grounding for moral standing.[ix] He claims that this cannot be so. You need to have some view on the ontological grounding for moral status in order to know which kinds of behavioural evidence might be relevant for assessing moral standing. For example, pain-related behaviours are only good evidence for moral standing if you already accept that sentience is the ontological grounding for moral status. This is correct and, as I conceded above, a proponent of a view like that of myself or Shevlin's cannot be strictly neutral with respect to the ontological grounding. That said, it is possible to be relatively neutral or quasi-agnostic with respect to ontological grounding. You can do this by arguing, as I did above, that if the ontological grounding is a mental or psychological one, then no matter which precise property or set of properties is thought to ground moral standing, behavioural equivalence will always provide evidence for its presence. Alternatively, and more boldly, you can argue that even if there is some doubt as to the exact set of mental properties that grounds moral status, behavioural equivalence of a diverse and consistent type should be enough to convince you that an entity possesses moral status ("I don't know exactly what it is that grounds moral status but this thing sure looks and acts consistently like other beings that have moral status so it probably does too").

Three final points are worth mentioning before discussing the relative merits of ethical behaviourism vis-à-vis cognitive equivalence. First, ethical behaviourism is a claim about the kinds of evidence that are sufficient for moral standing, not a claim about the kinds of evidence that are necessary for moral standing. In this respect, I agree with Shevlin that there could be other types of moral status beyond that of psychological patiency. For example, certain artifacts or natural wonders may have moral status that has nothing to do with their psychological properties. Second, ethical behaviourism does not deny that people can and do use non-behavioural evidence to make inferences to moral status. It just argues that these other forms of evidence are not as reliable and so do not defeat or undermine the behavioural evidence. This is something I discuss in a lot of detail in previous publications[x] and it is important to my analysis of Shevlin's views below: I would argue that evidence of cognitive equivalence is not as reliable nor as separable from behavioural evidence as he claims. Finally, although I have only focused on moral status in this discussion , I take ethical behaviourism to be a general strategy for deciding the ethical status of many relationships we might have with other beings. Thus, for example, I have argued that we can adopt an ethical behaviourist stance when trying to decide whether an artificial being is a friend[xi] or a lover.[xii] The only thing that varies across these different cases is the kinds of behaviours that an entity must display in order to count as a moral patient or a friend or a lover.

4. The Problem with Cognitive Equivalence

Shevlin, as noted in the introduction, sees some merit in my behaviourist strategy. He just thinks it has some significant limitations and that an alternative, cognitive equivalence strategy is superior. As best I can tell, the cognitive equivalence strategy would function almost exactly like the behavioural one. In fact, it may be intended to complement the behaviourist one in such a way that instead of focusing solely on behavioural evidence we should also focus on evidence relating to cognitive equivalence. In other words, behavioural evidence continues to count for something but it needs to be supplemented by cognitive evidence.

Specifically, in his paper, Shevlin suggests that when assessing moral status we should look to evidence of "episodic memory, metacognitive representation and affective states"[xiii] and also "working memory, a theory of mind…[the capacity for] 'mental time travel' and creative problem-solving", as well as the presence of an "internal representation system for registering 'desirable' and 'undesirable' events".[xiv] Evidence for the existence of these kinds of cognitive architecture and function are what we need to establish psychological moral patiency. That said, to be clear, Shevlin, like myself, is not prescriptive about the exact set of cognitive capacities and functions that establishes moral patiency. These examples are merely illustrative of the kind of thing that might be relevant.

Why does Shevlin think that the behaviourist strategy is insufficient and that the cognitive equivalence approach is superior? On my reading, he identifies three main problems with the behaviourist strategy.[xv]

The first is the *underinclusivity problem*. As noted above, the behavioural strategy relies heavily on comparisons and analogies. We work from entities whose moral status is already accepted to entities whose moral status is contested. This entails a kind of conservatism with respect to establishing claims of moral status. An entity has to be somewhat similar to another before the question of its moral status is entertained. What happens when you encounter an entity that shares no behavioural equivalency to humans or animals? As Shevlin points out, the behavioural strategy requires that, at best, you remain agnostic about its moral status.

But why is this a flaw with the behavioural strategy? Presumably, remaining agnostic about the moral status of such beings is only a problem if there is some good non-behavioural evidence for believing that the beings in question have moral status. In other words, it is only a problem if we think that the behaviourist approach is underinclusive with respect to claims of moral status. But the presence of good, non-behavioural evidence for moral status is exactly what the behavioural strategy calls into question. As I have defended it, this strategy claims that other kinds of evidence are unreliable or misleading.[xvi] So how can we know that the theory is underinclusive unless we implicitly reject its premise? To put it another way, this is scenario can only be a good criticism of the behavioural equivalence strategy if you have some other reason for thinking that behavioural evidence is insufficient. It doesn't work on its own.

The second problem is the *gaming problem.* If we apply the strategy in practice, we will, presumably have to specify that certain behavioural criteria provide good evidence of moral status. For example, we might stipulate that consistent approach and avoidance behaviours displayed by an artificial being provide good evidence for robust preferences and desires (this is illustrative only). Shevlin worries that if we specify such criteria there is a risk that they will be gamed. A sufficiently crafty AI engineer, for example, could design an AI that can perform these behaviours but do little else. We could then be tricked into thinking that an AI has moral status when it does not.

This is, of course, a criticism that has been directed at the Turing Test for machine intelligence for quite some time. It is a good critique in some ways. There is a danger that behavioural criteria will be gamed. That said, gaming is a problem for all tests that purport to establish the existence of a psychological or mental capacity, including Shevlin's own cognitive equivalence test (in fact, gaming might be a problem for all tests, period). The inner mental life of a being is never directly observable. We always have to work from outwardly visible signs and signals to inferences about inner mental life. Sometimes these signs and signals can be manipulated. The history of lie detection and guilty knowledge tests highlight this problem. Ostensibly, these tests work from the assumption that traditional behavioural tests for deceptiveness — such as listening to what someone has to say and trying to catch them out by asking tricky questions — can be gamed. They supplement these tests by looking to physiological and, more recently, neural markers of deceptiveness or guilty knowledge. But even physiological and neural marker-based tests can be gamed. The traditional lie detector test, for example, can be gamed through certain kinds of breathing and muscle clenching. Electroencephalography tests and fMRI tests can also be gamed through different concentration and attention-fixing practices. The bottom line is that gaming is certainly a problem but it is not clear that we can avoid it with any procedure for establishing moral status. The best we can do is to update the test parameters and insist on diverse and robust criteria for establishing behavioural (or, indeed, cognitive) equivalence.

The third problem that Shevlin discusses is probably the most important. Unlike the others there is no convenient shorthand name for this problem but, in lieu of a better one, I will call it the *'cognitive settler problem'*. The problem is this: there are tricky cases where behavioural evidence doesn't seem to settle moral status (or seems to be misleading); in these cases, we will need to appeal to some other kind of evidence. What evidence might that be? Shevlin argues that evidence regarding cognitive architecture will be crucial. An example of a tricky case is that of a hypothetical deep sea slug whose pain behaviour is similar to that of a more physiologically complex organism but who accomplishes this behaviour, not through a central nervous system, but through "peripheral mechanisms" in its skin.[xvii] The argument is that the absence of a more sophisticated cognitive architecture defeats the claim it might have to moral status. Other tricky cases include the sometimes sophisticated behaviours of decapitated and decerebrated animals and hypothetical alien creatures who don't look or act anything like humans or animals. The common variable across these cases is that the behavioural evidence points in one direction but we need to consider the evidence of cognitive equivalence before settling the issue of moral status.

These examples sound superficially plausible. I agree with Shevlin that most people would hesitate to believe that the pain behaviour of a sea slug is indicative of anything morally significant if it lacks a sophisticated cognitive architecture to support that behaviour. But the crucial question is not whether people might hesitate to do this but whether they are epistemically warranted in that hesitation? Shevlin seems to trust his intuitions about these cases more than I do. In my view, there are at least two problems with assuming that cognitive equivalence is what ought to settle these kinds of cases.

The first problem is that appealing to cognitive evidence looks an awful lot like begging the question. Why assume that the sea slug lacks moral status because it lacks a sophisticated cognitive architecture? Why assume that behaviour alone is not enough? The only obvious answer is because you are already precommitted to the cognitive equivalence approach. After all, you could just as easily endorse the moral status of the sea slug on the grounds of behavioural equivalence. In fact, I would go further and argue that we should do so on the grounds that behavioural evidence is the superior form of evidence for psychological patiency. Consider, for example, the case of people with congenital hydrocephalus. Their brains lack certain structures that most humans have (and there are other conditions with similar effects on the structure of the brain). Although the condition can be severe in some individuals, some can be relatively behaviourally normal. Should I disregard their moral status because they lack certain cortical structures that I think are 'normal'? I find it hard to accept that. Why adopt a different approach in the case of the sea slug? It can't be just because we think certain neural structures are more important than others because, as I will point out below, we can only reach that conclusion by using behavioural evidence. When confronted with the raw data that two entities behave in a similar way but one lacks a sophisticated cognitive architecture, we would have to be precommitted to the cognitive equivalence approach to suppose that this undermines a claim to moral status. In short, these kinds of tricky cases do not provide an argument for cognitive equivalence in and of themselves.

The second problem is more fundamental and may get to the heart of the disagreement between myself and Shevlin. The problem is that Shevlin seems to think that behavioural evidence and cognitive evidence are separable. I do not think that they are. After all, cognitive architectures do not speak for themselves. They speak through behaviour. The human cognitive architecture, for example, is not that differentiated at a biological level, particularly at the cortical level. You would be hard pressed to work out the cognitive function of different brain regions just by staring at MRI scans and microscopic slices of neural tissue. You need behavioural evidence to tell you what the cognitive architecture does. This is what has happened repeatedly in the history of neuro- and cognitive science. So, for example, we find that people with damage to particular regions of the brain exhibit some odd behaviours (lack of long term memory formation; irritability and impulsiveness; language deficits; and so on). We then use this behavioural evidence to build up a functional map of the cognitive architecture. If the map is detailed enough, someone might be able to infer certain psychological or mental states from patterns of activity in the cognitive architecture, but this is only because we first used behaviour to build up the functional map. Behavioural evidence remains the foundation. In any event, in practice, most tests for cognitive capacity are behavioural in nature. For example, we don't test for working memory or episodic memory by opening up people's skulls and seeing their internal cognitive architecture. We ask them questions, get them to perform tasks, and then use that behavioural evidence to infer the presence of some underlying cognitive architecture.

In sum, when we consider these tricky cases, I think we see that the cognitive equivalence strategy is not distinct from the behavioural equivalence strategy. They are one and the same thing. It's behaviour all the way down.

Or is it? I will say that there is something of a paradox at the heart of this debate. On the one hand, it seems like we need behaviour to make inferences about underlying cognitive capacity. On the other hand, it seems like we need to make assumptions about underlying cognitive capacity to make sense of behaviour. It is a bit of a bootstrapping paradox: how did this inferential loop get started? How did we first establish any link between cognition and behaviour? Did we just hypothesise or guess at some cognitive architecture based on behaviour? The most plausible answer, I suspect, is that we started by using our own internal experience, and its association with behaviour, as a guide. The problem, of course, is that we cannot rely on our own internal experience as evidence when it comes to the moral status of other beings.

Shevlin, Henry, (2020) "How could we know when a robot was a moral patient? Cambridge Quarterly. [– DETAILS TO BE SUPPLIED ONCE ISSUE AND PAGINATION IS SET.]

[i] Shevlin, Henry, "How could we know when a robot was a moral patient? Cambridge Quarterly 2020. [– DETAILS TO BE SUPPLIED ONCE ISSUE AND PAGINATION IS SET.]

[ii] Singer, Peter, *The Expanding Circle*. Princeton, NJ: Princeton University Press, 1981; Buchanan, A. and Powell, R., *The Evolution of Moral Progress: A Biocultural Theory*. Oxford, UK: OUP, 2018.

[iii] See, for example, Danaher, J., 'The Philosophical Case for Robot Friendship'. Journal of Posthuman Studies 2019, 3(1): 5-24; Danaher, J., 'Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism'. *Science and Engineering Ethics* 2020a, 26, 2023–2049; Danaher, J., 'Robot Betrayal: a guide to the ethics of robotic deception' *Ethics and Information Technology* 2020b, 22: 117–128; Danaher, J. 'Sexuality' In Dubber, Pasquale, Das (eds). Oxford Handbook of Ethics of AI. Oxford: Oxford University Press, 2020c

[iv] See note 3, Danaher 2020a.

[v] Nyholm, S., *Humans and robots: ethics, agency, and anthropomorphism*. London: Rowman Littlefield International, 2020.

[vi] Smids, J. 'Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot?' *Science and Engineering Ethics* 2020, https://doi.org/10.1007/s11948-020-00230-4

[vii] Nyholm actually focuses on other applications of ethical behaviourism in his criticism, particularly its application to friendship. This doesn't affect the point being made here though.

[viii] Cf Smids 2020, note 6, p 6 onwards.

[ix] See note 6, Smids 2020, p 10 onwards

[x] See note 3, Danaher 2020a.

[xi] See note 3, Danaher 2019

[xii] See note 3, Danaher 2020c

[xiii] See note 1, Shevlin 2020, **page number needed.**

[xiv] See note 1, Shevlin 2020, **page number needed.**

[xv] See note 1, Shevlin 2020, section 3.5

[xvi] See note 3, Danaher 2020a

[xvii] See note 1, Shevlin 2020, **page number needed**.