Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism

By John Danaher

(pre-publication draft of a paper forthcoming in *Science and Engineering Ethics*)

Abstract: Can robots have significant moral status? This is an emerging topic of debate among roboticists and ethicists. This paper makes three contributions to this debate. First, it presents a theory – 'ethical behaviourism' – which holds that robots can have significant moral status if they are *roughly performatively equivalent* to other entities that have significant moral status. This theory is then defended from seven objections. Second, taking this theoretical position onboard, it is argued that the performative threshold that robots need to cross in order to be afforded significant moral status may not be that high and that they may soon cross it (if they haven't done so already). Finally, the implications of this for our procreative duties to robots are considered, and it is argued that we may need to take seriously a duty of 'procreative beneficence' towards robots.

Keywords: Robots; Moral Status; Moral Standing; Ethical Behaviourism; Procreative Beneficence

Introduction

A debate has arisen about the moral and ethical status of robots (Gunkel 2018b). Do they or could they have *significant moral status*? This article makes three contributions to this debate. First, it presents and defends a theory —called 'ethical behaviourism' — which holds that robots can have significant moral status if they are *roughly performatively equivalent* to other entities that are commonly agreed to have significant moral status. An argument is presented in favour of this theory and it is then defended from seven objections. Second, taking this theory onboard, the article asks the obvious question: what kind of performative threshold must robots cross in order to be afforded significant moral status? Using analogies with entities to whom we already afford significant moral status, it is argued that the performative threshold may be quite low

and that robots may cross it soon (if not already). Third, and finally, the article considers the consequences of this for our procreative duties[1] to robots.

Some readers of this article may already accept the thesis it defends; some may be more skeptical. At the outset, it is important to speak to two of the potentially skeptical audiences. The first audience consists of those who think that the position defended is counterintuitive and absurd. Members of this audience are unlikely to be convinced by the end. But the goal is not to fully convince them. It is, instead, to open a dialogue with them and present a view that might encourage them to question their current theoretical commitments. The second audience consists in those who think that what is argued in this article is plausible but not particularly novel. Members of this audience will note that several authors have already defended the claim that we should take the moral status of robots seriously (e.g. Gunkel 2018a & 2018b; Coeckelbergh 2012; Sparrow 2004 & 2012; Levy 2009; Neely 2014; Schwitzgebel and Garza 2015). The debt to these authors is fully acknowledged. Where the present article differs from them is in (a) articulating a distinctive theoretical basis for this view; (b) defending this theory from a wide range of objections; and (c) pursuing more fully its practical consequences. The emphasis, consequently, should be less on the conclusions that are reached (though they are important), and more on the means by which they are reached. In this respect, the article should not be interpreted as defending a particular view as to whether robots *currently* can or should have moral status; rather, it should be interpreted as defending a particular view as to how we ought to resolve that question.

The Sophia Controversy and the Argument in Brief

To set up the argument it is worth considering a real-world controversy. Doing so illustrates that the issues addressed in this article are not of merely academic concern; they have immediate practical relevance. The controversy concerns one of the most widely-discussed social robots of recent years: the Sophia robot from Hanson Robotics. Sophia is visually human-like in 'her' facial appearance, gesture and voice. She has animatronics that enable her to recreate subtle gestures, including raised eyebrows,

[1] The term 'duty' is used in this paper in a sense that is interchangeable with cognate terms such as 'responsibility' or 'obligation'. It used to denote a normative requirement or restriction placed on someone's conduct that means that this conduct is not a matter of personal preference but is, rather, ethically mandated.

smiles and smirks. At the time of writing, her other human-like attributes, particularly conversation and locomotion, are more limited.[2]

Sophia has proved controversial among roboticists because of an incident that occurred on the 25th of October, 2017 at the Future Investment Initiative Conference in Riyadh, Saudi Arabia. In what many feel to be a marketing stunt, Sophia was granted Saudi Arabian citizenship (Stone 2017).[3] This led to a plethora of online criticism. For example, the AI-ethicist Joanna Bryson argued that the gesture was insulting given that the Saudi Arabian government does not recognise the full rights of many human beings, particularly women and migrant workers (Vincent 2017). Others argued that there was no rational justification for the move, given the current abilities of Sophia. Sarah Porter, founder of the World AI Summit, on January 10th 2018, underscored the absurdity of it all by saying that she was 'dressing [her] smartphone up in a dress, calling it Isobel and teaching it to walk' and wondering whether she could also, consequently, 'have coverage at all major tech news channels please?'.[4]

Comments like this pose a challenge for anyone claiming that robots could have significant moral status. They suggest that *performative artifice* by itself cannot suffice for moral status. Dressing a machine up in a human-like body, and making it look and act like a human, cannot be enough to ground its moral status. Something more is required, probably some internal mental apparatus (or 'soul') that enables the robot to feel, think, and see the world in a similar fashion to us. Writing in a different context, specifically in response to claims that we could have loving relationships with robots, Nyholm and Frank make this point by arguing that what 'goes on "on the inside" matters greatly' (2017, 223) when it comes to determining the ethical status of our relationships with robots.

This article rejects this view. It argues that what's going on 'on the inside' *does not matter* from an ethical perspective. Performative artifice, by itself, can be sufficient to ground a claim of moral status as long as the artifice results in *rough performative*

[2] To see what Sophia is like go to https://www.hansonrobotics.com/sophia/
[3] As Gunkel 2018b, p 116 points out, the gesture was not completely unprecedented. The Japanese have recognized a non-legal kind of robot citizenship in the past for artificial creatures such as Paro (the robotic seal).
[4] Porter's tweet is available at: https://twitter.com/SColesPorter/status/951042066561323008 (accessed 10/7/2018)

equivalency between a robot and another entity to whom we afford moral status. The argument works like this:[5]

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

The terminology requires clarification. If an entity has 'moral status' then it has moral standing and considerability (Jaworska and Tannenbaum 2018; Gruen 2017). Our treatment of that entity is not a matter of mere preference. There are ethical, not merely practical, limits to how we can treat it. If the entity has 'significant' moral status then those limits may be quite strict: we will not be allowed to mistreat or harm the entity without some overriding moral justification. Having moral status does not mean that the entity has legal *status* nor legal *rights*, at least not necessarily. There is a connection between legal status and moral status, but the practical justiciability of legal rights is often a relevant consideration when it comes to recognising those rights in another entity (Bryson et al 2017). Granting Sophia citizenship rights, for instance, is not required simply because we think she has significant moral status (if we do think that). This is important because the argument being defended in this article is not concerned with the legal rights of robots vis-a-vis humans but, rather, the ethical duties of humans vis-a-vis robots.

What kinds of ethical duties might humans owe to robots on the basis of this argument? This depends on the kinds of duties owed to the entities with whom robots are being performatively compared. The argument works off a principle of analogy: if

[5] The argument has some similarities with the 'no-relevant-difference' argument presented by Schwitzgebel and Garza (2015). But their argument is not grounded in the behaviourist view and is open to multiple possible understandings of a 'relevant difference'. It also encourages the search for disconfirming evidence over confirming evidence.

case A is like case B (in all important respects) then they should be treated alike. So, for example, if animals are owed certain moral duties – e.g. not to be mistreated or subjected to needless cruelty – and if robots are roughly performatively equivalent to animals, then, following this argument, robots are owed equivalent duties. What that means in practice depends on what cruelty or mistreatment consists in but, for example, there may be a duty not to physically damage a robot, or erase its memories, or switch it off without an overriding moral justification.

The other bit of terminology that needs clarification is that of 'rough performative equivalency'. This means that if a robot consistently behaves like another entity to whom we afford moral status, then it should be granted the same moral status. So 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. This is what it means to say that performative equivalency provides sufficient ground for equal moral status.

It is worth noting that the modifier 'rough' is included in recognition of the fact that no two entities with moral status are ever exactly performatively equivalent. For example, there will be a variety of performative differences between any two randomly-selected human beings: they will look slightly different, they will have different careers, different beliefs, different habits and so on. None of this means that they do not share significant moral status. There is enough rough equivalence for that property to be shared. This means that a robot need not look or behave *exactly* like another entity to whom we afford moral status in order for it to be afforded the same moral status. It is enough if it displays most of the relevant performative cues in similar circumstances. It is important to emphasise this point now because some people might see the inclusion of the modifier 'rough' in the motivating premise of the argument as an attempt to stack the deck in favour of the view that robots can have significant moral status. This is not the case. The idea that 'rough' performative equivalency is all that is required has independent validity.

That's enough by way of initial clarification. It is now time to defend the two main premises of the argument.

Defending Premise (1): The Case for Ethical Behaviourism

The case for premise (1) depends on a theory that is here called '*ethical behaviourism'*. Variations of this theory are hinted at in the writings of others (Sparrow 2004 & 2012; Levy 2009; Neely 2014) but it is believed that this article is the first to explicitly name it, and provide an extended defence of it.

To understand this theory, it is important to consider the similarities and dissimilarities between it and the classic methodological and ontological forms of behaviourism. Behaviourist psychologists like John Watson and BF Skinner favoured a *methodological* form of behaviourism. They thought it was scientifically improper for psychologists to postulate unobservable inner mental states to explain why humans and animals act the way they do. They felt that psychologists should concern themselves strictly with measurable, observable behavioural patterns (Graham 2015). Methodological behaviourism is what underlies the classic Turing Test for machine intelligence: Turing argued that we cannot observe the inner mental states that people think are constitutive of intelligence; all we can ever do is make inferences from observable behaviours (Turing 1950). As a methodological stance, behaviourism has much to recommend to it. Indeed, contemporary cognitive scientists, who are often said to have ditched behaviourism, are still behaviouristic in their *methods*. They still focus on recording and analysing external, measurable behaviour and brain phenomena, not inner mental states. They are just willing to hypothesise inner mental states to explain those external phenomena.

This methodological behaviourism should be contrasted with *ontological* behaviourism. Behaviourist philosophers, like Gilbert Ryle, once claimed that named mental states were really just abbreviations for sets of behaviours (Graham 2015). They argued that a statement like 'I believe X' was just a shorthand way of saying 'I will assert X in context Y', 'I will perform action A in pursuit of X in context Z' and so on. The mental, according to them, could be ontologically reduced to the behavioural.

Ethical behaviourism is an application of methodological behaviourism, not ontological behaviourism, to the ethical domain. Ethical behaviourism states that a sufficient *epistemic ground* or *warrant* for believing that we have duties and responsibilities toward other entities (or that they have rights against us) can be found in their observable behavioural relations and reactions to us (and to the world around them). It is the ethical equivalent of the Turing Test (Sparrow 2004 & 2012). It is a normative and epistemic thesis, not a metaphysical one. To be an ethical behaviourist one does not have to deny the existence of inner mental states, nor deny that those inner mental states provide the ultimate metaphysical ground for our ethical principles. Take consciousness/sentience as an example. Many people believe that humans and animals have moral status because they are sentient. An ethical behaviourist can accept this. They can agree that sentience provides the ultimate *metaphysical* warrant for our duties to animals and humans. They just then modify this by arguing that a sufficient epistemic warrant for believing in the existence of this metaphysical property can be derived from an entity's observable behavioural patterns. In other words, they will argue that a behaviourist epistemology constrains how we identify and apply the metaphysical properties relevant to moral status.

Why should one favour ethical behaviourism? The obvious reason is that it respects our epistemic limits. Although he may not agree with ethical behaviourism, [6] Kant provided one of the clearest articulations of these limits. He argued that we never have epistemic access to the metaphysical properties of the thing-in-itself; we only ever have access to its representations toward us. These representations may be used to infer the existence of certain metaphysical properties, but those properties cannot be epistemically grounded in direct contact with them: our contact with them is always mediated through representations. Ethical behaviourism argues that these limits carry over into practical ethics. Many principles concerning the moral status of others depend

[6] Kant was famously unwilling to accept that animals had moral status and drew a sharp distinction between practical reason (from which he derived his moral views) and theoretical reason (from which he derived his epistemological/metaphysical views). But others who have adopted a Kantian approach to philosophy have been more open to expanding the moral circle, e.g. Schopenhauer (on this see Puryear 2017). It is also worth noting, in passing, that the position adopted in the text has another affinity with Kantianism in that, just as Kant tended to reduce the metaphysical to the epistemological, ethical behaviourism tends to reduces the ethical to the epistemological. The author is indebted to an anonymous reviewer and Sven Nyholm for helping him to understand how Kant's reasoning relates to the argument defended in the text.

on metaphysical properties that cannot be directly assessed. For example, the most popular theories of moral status claim that it is because we think others are *conscious*, or have *high level cognitive capacities*, or are *persons* and have *interests*, that we owe them certain duties.[7] The ethical behaviourist points out that our ability to ascertain the existence of each and every one of these metaphysical properties is ultimately dependent on some inference from a set of behavioural representations. Behaviour is then, for practical purposes, the only insight we have into the metaphysical grounding for moral status.

The concept of 'behaviour' should be interpreted broadly. It is not limited to external physical behaviours (i.e.. the movement of limbs and lips); it includes all external observable patterns, including functional operations of the brain. This might seem contradictory, but it is not. Brain states are directly observable and recordable; mental states are not. Even in cognitive neuroscience few people think that observations of the brain are directly equivalent to observations of mental states. They may well infer correlations between those brain patterns and mental states, but they verify those correlations through other behavioural measures. For example, when a neuroscientist says that a particular pattern of brain activity correlates with the mental state of pleasure, they work this out by asking someone in a brain scanner what they are feeling when this pattern of activity is observed. They bootstrap from the behavioural to the mental to the neural. This primacy of the behavioural is often overlooked in popular conversations about cognitive neuroscience (Hare and Vincent 2016; Pardo & Patterson 2012; Bennett & Hacker 2003; Bennett, Dennett, Hacker & Searle 2007).

In short, then, the reason why one should accept ethical behaviourism is that it is an essential feature of day-to-day ethical practice: inferences from behaviour are the primary and most important source of knowledge about the moral status of others; if we did not rely on these inferences, the identification and protection of moral status would be impractical.

[7] For a comprehensive discussion of the potential metaphysical grounds for moral status, see Jaworska and Tannenbaum 2018. For specific discussions of consciousness, preference-satisfaction and personhood as grounds of moral status see Sebo 2018; Singer 2009; Regan 1983; and Warren 2000. For a discussion of the moral foundations of rights see Sumner 1987 and, as applied to robot rights, Gunkel 2018b.

To be clear, this is not an empirical thesis. The claim is not that everyone is, as a matter of fact, an ethical behaviourist. There are surely people who would disavow this view. It is, rather, a philosophical thesis. It claims that there are practical epistemic limits to how ethical principles can be applied. These limits apply whether people are aware of them or willing to acknowledge them. To put it another way, ethical behaviourism is a normative and meta-empirical thesis that: (a) tells us something about the kinds of empirical evidence that can be relied upon when thinking about moral status; and (b) how that evidence ought to be interpreted.

To make this more concrete, consider the following example. Philosophers of consciousness often talk about the possible existence of philosophical zombies (e.g. Chalmers 1996). These are entities that look and act like human beings but have none of the inner phenomenal conscious experiences of human beings. Assume, for the sake of argument, that the capacity to have phenomenally conscious experiences is the *sine qua non* of moral status. Then ask: how should we ethically treat a philosophical zombie? The ethical behaviourist answers that we should treat them the same as any ordinary human being. If a zombie looks and acts like an ordinary human being then there is no reason to think it does not share the same moral status. Ought implies can and, apart from the outward behavioural signs, there is no way to confirm or deny the presence of phenomenal states in others. So if phenomenal consciousness is to provide a practicable ground for moral status, it must be because it is cashed out in behavioural terms.[8] It is in this (epistemic) sense that what is going on "on the inside" does not matter from an ethical perspective. But this is an ethical conclusion only; nothing further is implied about the actual metaphysical nature of phenomenal consciousness.[9]

[8] An anonymous reviewer asks: what if it was a confirmed zombie? The ethical behaviourist would respond that this is an impossible hypothetical: one could not have confirmatory evidence of a kind that would suffice to undermine the behavioural evidence.

[9] One potential consequence of ethical behaviourism is that it should make us more skeptical of theories of moral status that purport to rely on highly uncertain or difficult to know properties. For example, some versions of sentientism hold an entity can be sentient without displaying any outward signs of sentience. But if this is correct, radical uncertainty about moral status might result since there is no behavioural evidence that could be pointed to that could confirm or disconfirm sentience. An ethical behaviourist would reject this approach to understanding sentience on the grounds that for sentience to work as a ground for moral status it would have to be knowable through some outward sign of sentience. For a longer discussion of sentience and moral uncertainty see Sebo 2018.

Ethical behaviourism has significant consequences when it comes to comparative assessments of moral status. An ethical behaviourist, when asked whether an entity (X) has moral rights and duties, knows that one easy way of determining this is to compare X's behavioural patterns to the patterns of another entity (Y) who already has some recognised moral status. If the two are behaviourally indistinguishable, the behaviourist will argue, in the interests of consistency, that X has those rights and duties too. In other words, a logical consequence of ethical behaviourism is that the following comparative principle ought to be applied to assessments of moral status:

> **The Comparative Principle of EB**: If an entity X displays or exhibits roughly equivalent behavioural patterns (P_1…P_n) to entity Y, and if it is believed that those patterns ground or justify our ascription of rights and duties to entity Y, then either (a) the same rights and duties must be ascribed to X or (b) the use of P_1…P_n to ground our ethical duties to Y must be reevaluated.

This 'performative equivalency' standard applies to debates about the moral status of robots. So if there is rough performative equivalence between a robot and another entity to whom moral duties are owed (where the equivalence relates specifically to the patterns that epistemically ground our duties to that other entity) it follows that the same duties are probably owed to the robot. This implies that performative artifice can, by itself, *suffice* for moral status.

It is not that straightforward, of course. There is a hedge in the comparative principle that suggests it can also be used to reevaluate the behavioural patterns used to ground our ethical beliefs. But that process of reevaluation confronts the same epistemic limits. Suppose it is agreed that duties are owed to animals due to their capacity to feel pain. The ethical behaviourist will argue that a sufficient epistemic ground for this belief lies in the observable behavioural repertoire of the animal, i.e. in the fact that it yelps or cries out when it is hurt, and recoils from certain pain-inducing objects in the world. Applying the comparative principle would imply that if a robot exhibits the same behavioural patterns, we owe it a similar set of duties. The use of those behavioural patterns to ground moral status could be reevaluated but ultimately any such reevaluation will result in another set of behaviourally-evidenced properties

being used to ground our ethical beliefs. At some point in time, people will have to settle on some set of behavioural patterns for grounding their beliefs about moral status, and, once they do, it will still be true that any entity that displays similar patterns of behaviour will warrant similar moral status. In other words: performative artifice is always sufficient for grounding moral status, even if there is some dispute about the precise contours of that performative artifice. This is gives us premise (1) of the argument: if ethical behaviourism is true, then robots that are roughly performatively equivalent to other entities that have significant moral status must be afforded that same status.

Many people will think this is wrong. To support their view, they might argue that other epistemically accessible facts play the critical role in grounding our beliefs about moral status. These facts constitute 'epistemic defeaters' to the performative equivalency standard. The remainder of this section looks at seven potential epistemic defeaters and argues that each fails to undermine the performative equivalency standard. The analysis of these defeaters is intended as a kind of 'proof by contradiction' for the ethical behaviourist approach (with the caveat that 'proof' is a strong word to use in ethics).

These objections are discussed with the obvious comparators of humans and (at least some) animals in mind. Although the inclusion of animals might be controversial, the idea that animals have *some* kind of moral status (one that means they can be harmed and their welfare needs to be considered in our decision-making about them) is widely accepted among moral philosophers and is respected in many legal systems. The objections are also discussed in order of generality, starting with those that take aim at the core idea of ethical behaviourism and continuting to objections that focus on behavioural anomalies that might undermine the application of the performative equivalency standard to robots.

Different ontologies objection

The first objection is that knowledge of ontology is what really matters when it comes to ascriptions of moral status. Humans and animals are biological beings, fashioned from complex assemblies of organic matter. Machines are non-biological

beings, fashioned from complex assemblies of inorganic matter. In other words, they are not made of the same stuff. Knowledge of this difference counts for something.

Arguments to this effect have featured in the abortion debate (e.g. Kaczor 2011). Opponents of abortion sometimes argue that being a member of the human species, or being a biological creature, is what determines the moral status of the foetus, not the functional/behavioural properties that pro-choice advocates favour. While this argument could have some relevance for the debate about robotic moral status, it is worth noting that it does not, by itself, contradict ethical behaviourism. Proponents of this argument are, presumably, claiming that species membership and/or biological properties are *metaphysical grounds* for granting moral status to a foetus; they are not necessarily denying that behavioural evidence can epistemically ground the ascription of such properties. It could well be that what determines membership of the human species or status as a biological being is the fact that an entity displays or exhibits certain behavioural tendencies and dispositions. Furthermore, even if that is wrong, no abortion opponent appears to reject the idea that an entity that displays the behaviour that is indicative of the properties ordinarily associated with moral status – e.g. consciousness, intelligence or personhood – should be denied moral status. In other words, they do not seem to deny that these things are *sufficient* for moral status. All they do argue is that species membership/biological status is an additional or independent ground for granting an entity moral status. This could be accepted, *arguendo*, for present purposes and it would not make a difference. The ethical behaviourist position advanced in this paper (as set out in premise 1) is that being performatively equivalent to an entity that already has moral status is *sufficient* for moral status. This does not mean that performative equivalency is *necessary* for moral status.

For the "ontology matters" objection to undermine the performative equivalency standard, its proponent will have to make the stronger claim that being made of the right stuff is necessary for moral status. But this is untenable because it results in an unjustifiable biological prejudice and mysterianism. To illustrate the point, suppose someone woke up one morning and was told that their spouse of the past twenty years is an alien from the Andromeda galaxy. The doctors have performed tests and it turns out that they have an entirely silicon-based biology. Nevertheless, they still act in the

same way, behave in the same way, and appear to be the same loving and supportive spouse that they always were (albeit with some explaining to do). Would the knowledge that they are made of different stuff warrant their being denied the moral status they have always been afforded? Or would the behavioural evidence negate the relevance of this new bit of information? The latter is the more plausible view than the former: it would require remarkable cruelty and indifference to their day-to-day interactions to reject their moral status. Similarly, suppose that (as seems to be increasingly possible) the biological parts of one's spouse were gradually replaced by functionally equivalent technological parts. After each and every replacement, they appear to be the same as they always were. At what level of cyborgisation should they lose moral status? Or should that happen at all? The most intuitively compelling view is that it shouldn't and hence that sharing a particular ontological essence is not necessary for moral status.

These are, admittedly, familiar thought experiments and ideas (cf. Schwitzgebel and Garza 2015). Nevertheless, they show that while knowing that an entity is made of biological/organic matter might (and the emphasis is on 'might') provide an additional sufficient criterion for ascribing moral status, it cannot undermine the independent sufficiency of the performative equivalency criterion.

Different efficient cause objection

A second objection is that the performative equivalency standard is undermined by our knowledge of different *efficient causes* of existence (i.e. of the different means through which an entity came into being) . It is known that animals and human beings have come into existence through a combination of evolution (which gives them their genetic constitution) and biological development (which shapes that genetic constitution into a specific form).[10] It is known that robots come into existence through a very different set of processes. They are programmed and manufactured by humans in labs or factories. Critics of performative equivalency might argue that knowledge of these different origins should block any inference from behaviour to moral status.

[10] A skeptic of evolution (e.g. a proponent of intelligent design) might dispute this, but if one believes in an intelligent designer then arguably one should perceive less of a morally significant difference between the efficient causes of humans and robots: both will be created by intelligent designers. That said, a theistic intelligent designer would have distinctive properties (omniscience, omnibenevolence) and those might make a difference to moral status. This issue is raised again in connection with the final cause objection, below.

That, at any rate, is what Michael Hauskeller appears to argue. Writing in the context of whether it is possible to have a loving relationship with a robot, Hauskeller initially seems to embrace a performative equivalency standard (Hauskeller 2017, 205), but then resiles from this by arguing that knowledge of the different efficient causes prevents us from concluding that a robot's behaviours have moral significance:

[A]s long as we have an alternative explanation for why [the robot] behaves that way (namely, that it has been designed and programmed to do so), we have no good reason to believe that its actions are expressive of anything [morally significant] at all. (Hauskeller 2017, 205)

There are, however, two reasons to reject this view. First, it is easy to overstate the differences between humans/animals and robots when it comes to their efficient causes. If the claim is that any entity that is designed and manufactured cannot have significant moral status, then we run into the problem that humans and animals can be designed and manufactured, at least in a certain sense, through careful planning and selective breeding. They are also likely to be susceptible to more invasive and precise forms of design and manufacture in the near future thanks to developments in genetic engineering. Someone might ethically oppose this kind of intervention into their developmental origin, but would they also thereby deny moral status to a being that is born as a result of them? This seems implausible: babies born as a result of genetic enhancement should not take a moral hit for the actions of their creators; they deserve the same moral status as any other children. Furthermore, it is worth nothing that evolution and biological development are themselves design and manufacturing processes that are not radically different from human design processes. Indeed, some programmers and manufacturers try to emulate the mutation and selection mechanisms of evolution, and the learning mechanisms of development, in their design of machines. Does this mean that machines that are created through such processes will have a moral status that they would otherwise lack? Again, it seems implausible. Behavioural criteria matter more.

Second, there is reason to think that origins (particularly biological origins) should not matter when it comes to determining our duties towards others. This is a more

controversial point, and not too much weight is rested on it here, but it could be argued that emerging norms concerning the treatment and status of, for example, transgender persons illustrate the relative unimportance of biological origins when it comes to ethical status. One interpretation of the literature on transgender rights is that we should not determine someone's ethically relevant status based on their biological origins but, rather, on how they authentically and consistently present themselves to us in everyday life**.** If this emerging norm is deemed morally appropriate, then it supports the performative equivalency standard and undermines the efficient cause objection.

Different final cause objection

There is, however, another way to interpret Hauskeller's concerns. Perhaps his worry is not about efficient causes and more about final causes? Maybe the concern is that robots will be designed to serve us or to serve the needs of their commercial or governmental proprietors, and that they will be owned and controlled by us or by these third party entities while serving these ends. All of these facts — which will be *known* in our interactions with the robots — will undermine the application of the performative equivalency standard to them.

In assessing this objection it is worth disentangling two things: (i) the ends that the robots have been designed to fulfil (serving us, supporting us etc.) and (ii) the social facts (ownership and control) that tend to be associated with beings that serve such ends (Bryson 2010 & 2018). If a robot is designed to fulfil a certain end, such as pleasing its owner, should this undermine its moral status? It is difficult to see why it should. The mere fact that an entity serves some end should not undercut its claim to moral status. If one adopts a naturalistic and evolutionary understanding of human origins, then one will more than likely accept that humans have been designed (by natural selection) to fulfil the ends of survival and reproduction. Even if humans rebel against these ends, they still lurk in the background and innate instincts and drives will push us toward those ends. But presumably this fact alone should not undermine human moral status. Likewise, if one adopts a theistic understanding of human origins, then one will more than likely accept that humans have been designed (by God) to fulfil certain ends, possibly including serving and worshipping Him. But, again, it seems highly unlikely

that one would take this to undermine a claim of human moral status.[11] The bottom line, then, is that just because an entity serves an end — up to and including an end that involves worship of another — it does not follow that its claim to moral status is undermined. Related to this, it is worth noting that certain robotic manufacturing processes — particularly those that incorporate machine learning — may result in robots that do not serve any clearly interpretible end or an end that is readily associated with their original creators. In this sense, modern robots may be much more like humans who have been loosely programmed by evolution and cultural development.

What about the fact that the entity is owned and controlled by another? This also should not undermine the performative equivalency standard. One reason for this is that the mere fact that an entity is owned or controlled does not, by itself, mean that the entity should not be treated with moral respect: this is reflected in many animal cruelty and welfare statutes.[12] Another reason is that, if anything, the fact of performative equivalency should cause us to reevaluate the system of ownership and control, not vice versa. Ownership and control are socially contingent facts. They are not baked-into the natural order. Not too long ago, humans owned and controlled other humans. This practice is no longer accepted because the full and equal moral status of the once-owned and controlled human beings is now recognised. Likewise, humans currently own and control many animals, but no one thinks this fact alone undermines claims to their moral status. If an animal rights activist argued that we ought to grant animals moral status it would be odd indeed if someone responded by arguing that this is not morally justified because animals are owned and controlled by humans. The animal rights activist would rightly argue that this is irrelevant. So too with robots. If robots are performatively equivalent to other entities to whom moral status is ascribed, then we may need to call into question the legal and institutional norms that (by default or

[11] If anything the opposite might be true. Theists might wish to ground moral status in non-observable metaphysical properties like the presence of a soul, but such properties run into the same problems as the strong form of sentientism (discussed in footnote 9). There are other possible religious approaches to moral status but religious believers confront similar epistemic limits to non-believers in the practical implementation of those approaches and this constrains how they can interpret and apply theories of moral status.

[12] Connected to this, an anonymous reviewer also points out that Kant (unlike many modern Kantians), in the *Metaphysics of Morals*, argued that although servitude was permissible servants were still owed duties of moral respect.

historical inertia) grant humans ownership and control over them. We should not use the fact of ownership and control to call into question the moral status of robots.

That said, the fact that humans once owned and controlled other humans – and, more generally, the fact that whole groups of humans were once systematically treated as having less than full moral status despite being performatively equivalent to humans that were recognised as having full moral status – might say something interesting about how humanity has thought about moral status in the past. It might suggest that the performative equivalency standard has not been our default historical approach to deciding questions of moral status.[13] But even if that is an accurate interpretation of the historical record it does not mean that the performative equivalency standard is morally unjustified in the future. On the contrary, the lesson of history could be that it was wrong to overlook performative equivalency in the past and it would be wrong to do so again.[14]

Deception and manipulation objection

A related objection — and possibly the one that captures many people's unease about Sophia and her alleged citizenship rights — is that any performative equivalency between robots and humans will be achieved through subterfuge, deception or manipulation. The manufacturers of robots will get their creations to mimic certain behavioural cues that we associate with beings with significant moral status, but because of the internal nature of the robots, these behavioural cues will not correlate with (or supervene upon) the metaphysical properties that we think ground moral status (e.g. conscious awareness and understanding). This seems to be what motivates the incredulity in the tweet cited earlier on and features heavily in Joanna Bryson's and other critics' arguments against the creation of person-like robots (Bryson 2010 and 2018; Leong and Selinger 2019).

[13] It might also be the case, as an anonymous reviewer points out, that our historical forebears conveniently overlooked or ignored the moral relevance of performative equivalency because doing so served other (e.g. economic) interests.

[14] It should also be noted that the historical mistreatment of groups of human beings would call into question other grounds of moral status such as ontology and efficient cause. So history does not speak against the performative equivalency standard any more than it speaks against those standards.

Fears about deception and subterfuge are, however, frequently misconstrued. Seeing why gets to the heart of what is distinctive about the ethical behaviourist stance. The behaviourist position is that even if one thinks that certain metaphysical states are the 'true' metaphysical basis for the ascription of moral status, one cannot get to them other than through the performative level. This means that if the entity one is concerned with consistently performs in ways that suggest that they feel pain (or have whatever property it is that we associate with moral status) then one cannot say that those performances are 'fake' or 'deceptive' merely because the entity is suspected of lacking some inner metaphysical essence that grounds the capacity to feel pain (or whatever). The performance itself can verify the presence of the metaphysical essence.

This does not mean that deception is impossible in the case of robots or that people cannot be deceived (or mistaken) as to the moral status of robots. The 'consistency' of the performance is critical here. Further behavioural examination or probing may reveal that the entity doesn't really doesn't really feel pain (or have whatever other property we care about). This may cause one to change one's opinion about their moral status. But this will only be because one has been exposed to countervailing behavioural evidence. To illustrate, imagine your friend showed up in your office one day, limping and whimpering about their sore leg. You suspect they are lying. How could you confirm this? Well, suppose you see them the next day running up and down the street, then jumping up and down on the spot one hundred times, and suppose you learn from conversation with others that they never said anything about feeling pain to anyone other than you. All that behavioural evidence (direct and indirect) would suggest that they were indeed faking it when they came into your office. It would give you countervailing behavioural warrant for disbelieving them. But if all the other behavioural evidence is consistent with their initial claims – if they limped up and down the street and complained to everyone about the pain – you would not have warrant for disbelieving them. Either way, it is the presence or absence of consistent behavioural performances that determines whether they are deceiving you or faking it; it is not the presence or absence of some inner metaphysical essence. That is not something that can be observed and verified. The same approach should apply to our interactions with robots.

Relatedly, it could well be the case that some people are lazy or cognitively impaired and thus unable to do a proper consistency check. Those people may leap to unwarranted conclusions about the moral status of robots. This should be avoided: equivalency must be properly tested. Nevertheless, it is always the behavioural evidence that determines whether the judgment about moral status is warranted, not some mismatch between the behaviour and some other internal factor.

There is a modified version of this objection that might seem more persuasive. Someone could accept the behaviourist stance and yet argue that robotic performances are fake because they do not emanate from a mechanism that is functionally equivalent to the human brain. In other words, they could argue that behavioural states can only verifiably ground ascriptions of moral status if they are correlated with the right kinds of functional brain state. If those functional brain states are not present, then there is deception or fakery.

There are two problems with this. First, the earlier comments about the need for an expansive interpretation of the word 'behavioural' should be remembered. If you really believe that these are relevant to the ascription of moral status then they can be included within the performative equivalency standard that robots would have to match (Raoult and Yampolskiy 2018). This would still imply that it is possible for robots to be granted moral status on the basis of performative equivalency, provided that we don't commit ourselves to the implausible version of biological mysterianism that was outlined earlier. Second, notwithstanding this possibility, it is probably wrong to think that the presence of functional brain states really does make a critical moral difference. For one thing, our understanding of the relationship between brain states and morally significant metaphysical states (such as sentience and personhood) is pretty iffy. There doesn't seem to be any reason to think that specific functional brain states are the only way in which to realise those metaphysical states. Furthermore, as argued earlier, our epistemic warrant for associating functional brain states with metaphysically significant properties like sentience is ultimately verified by behavioural evidence. This evidence consequently should have epistemic primacy in our ethical practices.

The 'Thinking Otherwise' objection

Another objection to the performative equivalency standard might be found in the work of Coeckelbergh and Gunkel (2014 and 2016). Coeckelbergh and Gunkel argue against the 'properties approach' to determining moral status in both animal and machine ethics (Gunkel 2018a and 2018b). The properties approach holds that whether or not an entity deserves moral status depends on whether it exemplifies certain properties, such as the capacity for suffering or the capacity to be the subject of a life. The properties approach is similar to the approach advocated in this article, with the caveat that this article claims that behaviour provides sufficient warrant for believing in the existence of such properties.

Gunkel and Coeckelbergh present four criticisms of the properties approach. They argue that it proceeds from an unexamined *anthropocentric bias*: proponents of the approach start with properties that humans exemplify, such as sentience or self-awareness, and then work outwards from those properties to determine the moral status of others. They argue that the approach is beset by *epistemological problems*: many of the properties favoured are epistemically opaque and it is not clear how we could know whether or not they are present. They argue that the approach creates *an illusion of neutrality* when it comes to determining moral status: the assumption is that the presence or absence of the relevant properties can be objectively and neutrally determined, and these are matters to be determined by scientists and animal behaviourists, not ethicists, but this ignores how deeply moral/ethical the determination of moral status really is. And finally, they argue that the properties approach often involves *sticking with a traditional and defective method* for determining moral status: the decisions as to which properties 'count' are ones that are made before people are born and are deeply embedded in contingent social norms and practices. This is why, historically, women and slaves were excluded from having moral status. To persist with the properties approach is to persist with these dubious social and cultural norms.

There is merit to each of these criticisms but they do not undermine the performative equivalency standard. The second criticism, relating to the epistemic opacity of properties, is the bullet that the ethical behaviourist thinks everyone should bite, and the third and fourth criticisms are consistent with performative equivalency.

All they really do is give reason to adopt a looser 'ethically precautionary' standard of performative equivalency in order to widen the scope of moral status and avoid status quo bias (this is discussed in more detail below). Finally, the first criticism highlights something that is arguably unavoidable in this arena. It is difficult to see how else we can proceed with ascriptions of moral status except outwards from what is known about humans. This doesn't bind us to a human-likeness standard of performative equivalency; but it does mean that human-likeness (at a minimum) should be sufficient for moral status.

Gunkel and Coeckelbergh go beyond critique and argue for an alternative 'relational' approach to moral status. This relational approach focuses on how other beings relate to us and enter into our lives. They base this on the work of the phenomenologist Emmanuel Levinas. He argued that the primary fact of existence was its relationality, i.e. the fact that we are in the world with others who intrude upon us in various ways. This intrusion necessitates a moral response, and as part of that response our relations with others must be parsed into ontological categories where some entities are seen to 'take on a face' and require special treatment. This 'taking on a face' is equivalent to acquiring moral status and entering a moral community. Coeckelbergh and Gunkel ask what it takes for an animal or robot (or 'Other') to take on a face. They think that asking this question takes us away from the properties-oriented mindset and focuses instead on our embodied interpersonal interactions with the Other. Discussing animals,[15] Coeckelbergh and Gunkel single out two things that seem to be quite important in determining whether animals take on a face. The first is the 'naming' of the animal: Giving an animal a proper name is a speech act with moral consequences. It draws the animal inside the moral circle. The second is the physical location of the animal: Animals that live outside our homes — in the fields and countryside — are different from animals that share our homes. By inviting them into our homes we also invite them into our moral circles (Coeckelbergh and Gunkel, 2014, 727).

The relational approach is provocative but also does not undermine the performative equivalency standard. On the contrary, the relational approach is consistent with the behaviourist approach. Both argue that the actual metaphysics of

[15] Gunkel has subsequently (2018a and 2018b) expanded the analysis to include robots

animals and robots is an ethical distraction. The focus should instead be on how they represent themselves to us. When it comes to the practical application of the relational approach, all that Gunkel and Coeckelbergh do differently is suggest that additional representational criteria, beyond the performative, and including the social-relational, might be relevant for ascriptions of moral status. Adding these criteria to the moral calculus could well be warranted, but then the sufficiency-necessity debate rears its head again. These other criteria might be sufficient (when taken with other factors) for an ascription of moral status, but are they necessary? If one had a robot that was performatively equivalent to a human should the fact that it did not have a name, or did not enter into embodied relations with humans, make a critical difference? It is hard to see why it should. It is more plausible to suggest that performative equivalency would trump these other considerations.[16]

In short, the relational approach is not opposed to the behaviourist approach. That said, there is, one critical difference. Gunkel and Coeckelbergh claim the relational approach does not give us clear ethical guidance on how to treat animals and machines, nor is it intended to (Coeckelbergh and Gunkel, 2014, 730; Gunkel 2018a, 95ff). The ethical behaviourist approach does: It says moral status should be granted to an entity when it is performatively equivalent to another entity to whom it is already granted.

(f) The Ontological and Practical Weirdness of Robotic Moral Status

The sixth objection holds that robotic moral status cannot be recognised solely on the grounds of performative equivalency because it would be too weird or unusual if it were recognised. Even if robots look and act like humans and animals they are still, underneath it all, very different. Their embodied parts can be easily replaced in the event of injury or accident; their 'minds' and memories can be backed-up and restored after 'fatal' error. They are not as fragile or morally needy as humans or animals. Their existential robustness means that, even if they are performatively equivalent to humans or animals, they don't need to be morally respected in the same way.

[16] That said, the performative equivalency view is not necessarily in tension with the relational view because the fact that people want to give robots names, invite them into their homes, and make them human-like in other ways is probably what drives people to create robots that are performatively equivalent. The author is indebted to an anonymous reviewer for suggesting this point.

This is a variant on the ontological differences objection considered earlier, albeit one focused on the ontological weirdness of robots. It is a common objection.[17] While it is superficially attractive, it has some problems. For one thing, just because a robot's moral status is recognised on the grounds of performative equivalency it does not follow that they need the exact same protections as the entity to whom they are being compared. Two humans can be roughly performatively equivalent and yet not be owed the exact same protections. This is something that needs to be worked out. More importantly though, if one believes that ontological robustness – specifically the fungibility/replaceability of functional parts –undermines claims to moral status, one starts down a very slippery slope (Carter and Palermos 2016). After all, human biological parts are increasingly replaceable with either organ transplants or artificial analogues to biological organs. Does the fact that one could give someone a cochlear implant make it okay to induce deafness? Surely not and surely robots that are performatively equivalent to humans should be afforded the same moral respect.[18]

Cumulative difference objection

The final objection is that although none of the preceding differences suffices to undermine the performative equivalency standard by themselves, taken as a collective they do. In other words, because robots and humans/animals differ in so many ways (biological constitution, efficient cause, final cause, etc) they cannot have the same moral status merely because they are performatively equivalent. One or two differences could be tolerated, but not so many.

This is tempting but it is difficult to fashion a morally significant difference out of a collection of differences that are, if the preceding rebuttals are correct, individually morally irrelevant (with the possible exception of species membership, which was only granted *arguendo*). It is at least deeply mysterious as to how this could happen. The burden of proof should be on the proponent of the cumulative difference argument to come up with a mechanism that allows for this.

[17] Schwitzgebel and Garza (2015) have an extended discussion of AI-fragility (or the lack thereof) and what it might mean for moral status. They initially agree with the position adopted in this article but also suggest that certain aspects of machine ontology might warrant greater moral protection.

[18] Contrariwise, if replaceability undermines the need for certain kinds of moral protections, then perhaps we need a new set of moral norms for entities that are easily replaceable. But this new set of norms would then apply to humans just as much as it would apply to robots.

One way that the argument might work is if it is assumed that the preceding arguments have not shown that these other factors are morally irrelevant but, rather, that they carry less weight than might initially be thought. One might then argue that, taken together, all these differences amount to something weighty. But then the critical question is whether the combined weight would be sufficient to defeat the inference from performative equivalency. This seems implausible. Performative equivalency would swamp everything else: to deny moral status to a robot that acted like a human in every important respect, but was made of different stuff and came into being in a different way, would be unduly reckless and insensitive. Performative equivalency is a decisive factor when it comes to ascriptions of moral status. It is sufficient for moral status.

Defending Premise (2): What's the performative threshold?

To this point, the focus has been on premise (1) and how the performative equivalency standard works in the abstract. Little has been said about premise (2) and how that standard could work in practice. That's where the spotlight now shifts.

When thinking about the practical application of the standard, it is important to emphasise that the argument defended in this paper is *strictly agnostic* when it comes to the precise content of the performative equivalency standard and the metaphysical assumptions that undergird it. Indeed, for the argument to work, there need not be any agreement on the metaphysical basis for moral status. As long as it is thought that some beings have moral status, and as long as the performative comparisons with other beings are roughly equivalent, ethical behaviourism kicks-in and logical consistency demands an ascription of moral status. This is because, to reiterate, ethical behaviourism is a meta-empirical thesis about how we ought to interpret empirical evidence concerning behaviour, not an empirical or metaphysical thesis about the kinds of behaviour we should be focusing our attention on. Nevertheless, whether the performative standard should be set at a 'high' or 'low' level and whether robots could meet that standard is something that can considered in this article. So although it may not be possible to reach a final conclusion as to what the performative standard should be – that would require a separate analysis – but it should at least be possible to sketch

some of the difficult compromises and tradeoffs that arise when choosing between a high or low level standard.

A 'high' level standard would require a behaviourally-sophisticated robot that is perfomatively equivalent to a competent adult human. It might be very difficult, but probably not impossible, for a robot to be created that satisfies this standard at some point in the future. There would, of course, be debates to be had about which adult human behaviours would be most critical and would have to be mimicked by the robot. It would be odd if we required the robots to look and act *exactly* like an adult human. Some of the things that adult humans do are not necessary for moral status. For example, if the robot doesn't fidget or scratch its nose or sweat, it would be odd to deny its moral status if it is otherwise performatively equivalent. Similarly, though this might be more controversial, full human-likeness in appearance (human-like skin, bipedality, facial gestures) would not seem necessary for moral status. In this respect the analogy between a high performative threshold and the Turing Test for intelligence might be quite strong: they might just be the same test on the grounds that it is cognitive behaviour that really matters when it comes to moral status.

What is more important for present purposes is whether setting the standard at a high level is justified. It might be *prudentially* justified. Given the controversy around Sophia's citizenship and the deep concerns expressed about the idea of robot rights in some quarters (e.g Bryson et al 2017; Bryson 2018), and given the fact humans are quick to anthropomorphise and over-ascribe agency to non-living things, it might avoid a lot of anger and strife if we set the standard at a high level. It would reassure the skeptics and naysayers that the day when we must be wary of our duties to robots is a long way off, and might serve as a check on our natural cognitive bias toward anthropomorphism. One might even argue that this stance is morally justified on the grounds that the differences between robots and humans discussed in the previous section, though morally insignificant in their own right, are nevertheless sources of normative uncertainty and, given the consequences of recognising robotic moral status for other ethical duties (e.g. the procreative duties discussed later on), one should err on the side of under-inclusivity when it comes to such uncertainty.

The problem with this argument is that normative uncertainty regarding moral status is usually thought to warrant over-inclusivity rather than under-inclusivity (Lockhart 2000; Guerrero 2007; Moller 2011; Neely 2014; Sebo 2018). Several philosophers have argued that normative uncertainty regarding the status of animals and foetuses should cause us to err on the side of including them within the circle of moral concern, not excluding them (Moller 2011; Sebo 2018). Erica Neely (2014) has defended this same view with respect to robots. The thinking is that the moral risk attached to over-inclusivity is much lower than the moral risk attached to under-inclusivity. It is worse to exclude a deserving entity from the circle of moral concern than to include an undeserving entity.

This might push us to favour a low-level standard, but if a low-level standard is favoured we run into another problem, namely: that robots may already be pretty close to meeting it, if they haven't done so already. This can be seen by examining some comparator cases. These may prove controversial, but the controversy can be addressed after they have been set out. The first comparator case is: persons with severe, permanent cognitive and/or physical disabilities. It is already commonly agreed that such persons have moral status and are owed duties despite the fact that their behavioural repertoires are limited. This applies to persons who are severely disabled from birth as well as people who become severely disabled later in life. The second comparator is: animals with minimal behavioural repertoires (e.g. chickens or mice). Although more controversial, many people now accept that such animals have *some* moral status, at least one that requires us to have concern for their welfare and not to subject them to unnecessary violence or harm without good cause. It doesn't seem too much of a stretch to suggest that robots could, if they are not already at least in the very near future, be performatively equivalent to one or both of these groups. The normative preference for over-inclusivity could then kick-in and justify affording them moral status.

This can be further underlined by considering in more detail the case of humans with severe cognitive and/or physical disabilities. Disability rights activists have argued that the performative threshold that such people have to meet in order to have their actions protected under Article 12 of the UN Convention on the Rights of Persons with

Disabilities[19] should be set at a low level. Arstein-Kerslake and Flynn (2017), for example, have argued that once the legal personhood of an individual with a disability is accepted, any action that they perform (with legal consequences) that provides evidence of 'intention' should trigger legal protection. In saying this, they both (a) explicitly acknowledge that the behavioural evidence of intention could be quite minimal and (b) that we should err on the side of over-inclusivity:

> "*We propose that…any indication that there was purpose and deliberation behind a particular action, decision, or omission, should be considered sufficient evidence to ascribe intention… If there is doubt about whether or not intention exists in an action taken by a person with a disability, we propose that, for the purposes of Article 12, an assumption is made in favor of finding intention…it is more dangerous to deny moral agency to people with disabilities than it is to simply accept that all people have moral agency and to then explore how best to [protect] the expression of that agency.*"
>
> (Arstein-Kerslake and Flynn, 2017, 26 - references omitted)

This analogy with persons with severe disabilities may provoke a negative reaction. Critics might argue that although it seems progressive to 'expand the circle' of moral concern to include robots (Singer 1981), doing so using this analogy insults the struggle of a historically oppressed and ignored group of persons who fought hard to get included in the circle of moral concern.[20] They could also argue that it risks underestimating the actual behavioural capacities of these individuals.

We should be to these concerns and they may provide a strong prudential reason to worry about favouring a low threshold. But it is important to probe the consequences of our moral practices and beliefs, even if this sometimes gets us into uncomfortable territory. If the performative equivalency standard is correct, and if we should err on the side of over-inclusivity when it comes to moral status, then the possibility of a low

[19] Article 12 of the UNCRPD recognises the right to equal recognition before the law of persons with disabilities and includes, specifically, the right to recognition of legal capacity (roughly: the capacity to make decisions on their own behalf).

[20] For an example of how this objection might play out, consider the controversy that Rebecca Tuvel's article 'In Defence of Transracialism' (2017) provoked when she argued that transgenderism and transracialism could be viewed as analogous phenomena.

threshold has to be accepted. It should also be noted that, if there is some worry about the insensitivity of the analogy with persons with severe disabilities, the analogy with animals is likely to prove much less offensive and the argument can be sustained just with that analogy. Finally, it should be remembered that even the most sophisticated of present day robots may fail to cross the low performative threshold set by these comparators. This is something that has to be carefully determined by detailed inquiry, not simply asserted or hypothesised in the abstract.

That said, there may be some moral reasons for rejecting a low threshold in the case of robots, but accepting it in the case of animals and persons with severe disabilities. It could be that other criteria for moral status are satisfied in these cases and this justifies a lower performative threshold. Perhaps, for example, if an entity has the *potential* for displaying a more sophisticated behavioural repertoire at a later time, it is okay to ascribe moral status for lower level performances at another point in time. Potentiality principles of this sort have long been used in the abortion debate to make the case for foetal moral status. The problem with using them in this context is that it does not seem right to say that animals or persons with severe permanent disabilities have the *potential* for higher level performances. At least, they do not have a very robust potential for such performances and it would be possible to argue that robots have this less robust potentiality too (if they just had the right technological modifications…). More plausibly, it could be argued that animals and persons with severe disabilities warrant a lower threshold because they are biological beings or members of the human species. This was rejected as a reason for denying the performative equivalency standard above, but it was conceded, *arguendo,* that it might provide an additional ground for moral status. That concession could be cashed in here and it could be argued that persons with severe disabilities and/or animals warrant a low performative threshold because they satisfy this additional criterion for moral status. This means that ontology must makes some moral difference (in this case by lowering the performative threshold required for moral status), and accepting this will have important, and possibly unpalatable, knock-on effects on other moral beliefs (such as our approach to abortion rights). There are also some other less defensible reasons to favour a low threshold. For example, one could cling to the status quo for ineffable/conservative reasons. Or, one could accept the analogy between robots and these other groups but

use this to call into question the low performative thresholds that are applied to animals and persons with severe cognitive and physical disabilities. Either response would have significant, probably unwelcome, repercussions.

In the end, neither a high level nor a low level performative threshold seems to be entirely desirable. If a high threshold is favoured, it would imply that robots are unlikely be granted significant moral status any time soon, and while that might be prudentially defensible it does not seem to be morally defensible. If a low-level threshold is favoured, this could imply that robots already have (or are very close to having) significant moral status. This would force a reevaluation of the morality of our actions toward them. Intermediate options, such as the 'low threshold + some additional criterion', or a moderate threshold that is defined independently, would suffer from similar problems. It seems that no matter what standard is picked, there are important tradeoffs to consider. Bringing robots within the circle of moral concern, or leaving them out, will have knock-on effects on our other moral beliefs and practices. These need to be carefully mapped, identified and evaluated.

Procreative Beneficence and Robots

If the argument to this point is successful, it follows that determining whether or not a robot has significant moral status depends on whether that robot is performatively equivalent to another entity with significant moral status. Assuming it is possible for us to create such a robot, it would then seem, prima facie, to follow that this is a decision with serious ethical ramifications. Might there be a duty to create or refrain from creating such beings?

Joanna Bryson is one of the few to appreciate the importance of this question. Bryson believes it is *possible* for us to create robots with significant moral status (Bryson 2010 and 2018). She is, however, adamant that we shoud not do this because there are too many costs. It would mean: that robots could not be used for beneficial instrumental purposes; that humans could avoid legal and moral responsibility for the actions of the robots they create; and that our current normative equilibrium would be destabilised (Bryson 2018). She argues that this does not mean that robots should be banned altogether; it just means that we should refrain from creating ones that attract

moral status. For the most part, Bryson focuses on the costs that creating robots with moral status would impose on human beings, but she also considers the negative impact this might have on robots too:

"..the policies I promote [reference omitted] have always explicitly considered the welfare of potential intelligent artefacts [i.e. robots]…Why should we design [such] artefacts to be in the position of competing with us for resources; of longing for higher social status (as all evolved social vertebrates do); of fearing injury, extinction, or humiliation?" (Bryson 2018, 8-9)

Of course, this reasoning would apply equally well to the procreation of animals and human offspring, so it is not clear how seriously it should be taken[21] and, as Danaher (2018) argues, there are potential benefits to the creation of robotic offspring that could outweigh these costs.

Nevertheless, Bryson's analysis is insightful because it suggests that the creation of robots with significant moral status can be viewed through the lens of procreative ethics. This is helpful because the literature on procreative ethics has identified many principles and constraints that ought to apply to procreative decision-making and it would be worthwhile seeing how (if at all) these carry over to the robot case. This article concludes by considering one obvious principle that might apply: the principle of procreative beneficence (PPB).

The PPB was originally formulated by Julian Savulescu (2001) and holds that although one is not under any obligation to procreate, if one decides to procreate one is under a duty to procreate a child with the best possible life given current knowledge and technology (Savulescu 2001, 415). The PPB is controversial when applied to human procreation (Holland 2016; Saunders 2015 & 2016; Overall 2011). It is criticised on three main grounds. The first is that Savulescu favours a welfarist test for the 'best life' that ignores or overlooks other aspects of the good life; the second is that there is no easy way to identify the child that will have the best life at the point of procreation

[21] Bryson may think there are other moral/ethical benefits that outweigh these costs in the case of humans — it is not clear from her writings. What is clear is that she thinks that human well-being trumps robotic well-being.

(Savulescu's preferred method of pre-implantation genetic diagnosis has its epistemic limits); and the third is that the PPB places too high a burden on potential procreators, particularly women.[22] Thus, the tendency is to think that, as applied to human procreation, the PPB, at best, identifies something that is morally supererogatory; and, at worst, not defensible at all.

But how does it fare when applied to robotic procreation? Contrary to what one might think, it may be less controversial in that case than in the human case. The reason for this is that the most compelling objections to the application of the PPB to human procreation — the excessive burden argument, and the epistemic limitation argument — carry much less weight when applied to robotic procreation. Consider Christine Overall's criticisms of the PPB. She argues that Savulescu unduly ignores the practical impact that the PPB would have on women. If followed, it would require women to forego the good of conception via sexual intercourse, and instead opt for IVF combined with pre-implantation genetic diagnosis, which is both expensive and carries significant medical risks (e.g. multiple gestations). It would, as Overall puts it, be forcing women to conduct a 'massive medical experiment' on their own bodies and that of their children for uncertain gain (Overall 2011, 127).

Similar concerns do not arise in the case of robotic procreation. Requiring robot manufacturers to create robots with the best possible life will undoubtedly impose burdens on them, but these burdens are not unreasonable. The decision to create a robot is entirely voluntary, and ensuring that, of the possible ones that could be created, creating the robot with the best possible life (given current technological limits), will not require one to forgo other decisive/overwhelming goods, or result in a problematically gendered distribution of risk and reward. Furthermore, many of the epistemic limitations that apply to human procreation would not apply to robotic procreation. The link between genetic constitution and the overall quality of life is, still, relatively uncertain. There are some genetic endowments that carry significant health risks, but beyond clearcut cases the ability to ensure the best possible life from genetic

[22] It is also sometimes criticized for being redolent of eugenics. However, Savulescu would argue that there are significant moral differences between what he is proposing and the morally repugnant policies of historical eugenicists. He is not claiming that people ought to be sterilized or prevented from having children in the interests of racial or cognitive purity. He is focusing on the need to benefit potential offspring; not on harming or restricting potential parents.

testing alone is limited. Controlling a robot's programming and technical constitution will be much more feasible. There can be more fine-grained control over its quality of life, though there will be some limitations (e.g. hacking, unforeseen social or natural disasters).[23]

There are two obvious objections to the use of the PPB in the case of robots. The first is that there is no meaningful concept of well-being or welfare that can be used to assess a robot's quality of life. Hence, the principle cannot be applied. This objection can be quickly dispatched. The key lesson of ethical behaviourism is that determining whether an entity's life is going well or going badly can be sufficiently assessed using behavioural criteria. In determining the welfare of humans, there is a tendency to focus on questions like: Are they learning? Do they have friends? Are they physically fit and able? Do they have a sense of purpose? Are their desires being fulfilled? None of these things can be determined by direct testing of an individual's metaphysical constitution. All of it must be determined by reference to the individual's performances and representations. A robot's quality of life can be assessed in a similar fashion, again using a performative equivalency standard.

The second objection is that there is at least one important difference between human procreation and robot procreation. If one decides to procreate a human being, then one has no choice but to procreate an entity with significant moral status. Human infants are, by necessity, beings with such status. If one decides to procreate a robot, then one does have a choice. One could choose to create a robot that lacks significant moral status (that fails to meet the performative threshold). This, in fact, is one of Bryson's main points when she argues that we should avoid creating person-like robots.

There are, however, two important limitations to Bryson's strategy. The first is that the drive to create robots that cross the performative threshold (which, as noted above, could be quite low) will probably prove too overwhelming for any system of norms (legal or moral) to constrain. And once the first robot crosses the performative

[23] Matthijs Maas has suggested to the present author that the hacking risk is quite severe. As he sees it "if a robot capable of suffering gets hacked, this would allow the attacker to inflict massively scalable, unbounded suffering or indignity on the AI (e.g. by speeding up its clock-time, making it suffer subjective millennia of humiliation). The amount of suffering that could be inflicted on a robot is therefore much higher than that which could be inflicted on a human". This might give very good reason not to create robots with moral status.

threshold, the PPB will have to be given some consideration because robots with moral status and better lives will be a 'live' possibility. The other limitation is more important. It might be very difficult to create robots that do not have some significant moral status, particularly if the performative threshold for moral status is low. It may require creating a robot that lacks any behavioural manifestation of intention, desire, or agency, which seems tantamount to requiring that robots not be created at all. As Gunkel notes (2018a, 94), following Bryson's strategy would require a form of robot asceticism which would be very difficult to police and maintain in practice.

None of this is to say that there is a general duty to create robots with significant moral status. The decision to do so – like the decision to procreate human offspring – is voluntary and subject to other moral constraints. If the resources to care for them are absent, or if it would compromise the well-being of other morally significant, and already existent, beings, then maybe they should not be created. But *if* there is some attempt to do so, it is not implausible to suggest that this decision should be constrained by the PPB.

Conclusion

In conclusion, this article has argued that performative artifice, by itself, can suffice for a claim of robotic moral status. If a robot looks and acts like a being to whom moral status is afforded then it should be afforded the same moral status, irrespective of what it is made from or how it was designed/manufactured. This is a counterintuitive view, but hopefully it has been shown to be defensible and worthy of consideration. If accepted, it has significant consequences not only for how robots that have already come into being are treated, but also for the principles by which decisions are made to create those robots in the first place. Although it is tempting to think that robots would have to demonstrate a high level of performative equivalency before they should be afforded moral status, there are good moral reasons to think that this is not the case. Furthermore, it could be that, because robots can acquire moral status through performative artifice alone, the decision to create them is governed by the principle of procreative beneficence.

Acknowledgements: The author would like to thank Matthijs Maas, Sven Nyholm and four anonymous reviewers for helpful comments on earlier drafts of this article. He would also like to thank audiences at NUI Galway and Manchester University for enduring earlier presentations of its core argument.

References

Arstein-Kerslake, A., & Flynn, E. (2017). The right to legal agency: Domination, disability and the protections of Article 12 of the Convention on the Rights of Persons with Disabilities. *International Journal of Law in Context*, 13(1): 22-38

Bennett, M.R. and Hacker, P.M.S. (2003). *Philosophical Foundations of Neuroscience*. Oxford, UK: Blackwell Publishing.

Bennet, M.R., Dennett, D., Hacker, P.M.S., and Searle, J (2007). *Neuroscience and Philosophy: Brain, Mind, and Language*. Oxford, UK: Blackwell Publishing.

Bryson, J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psycho- logical, ethical and design issues* (pp. 63–74). Amsterdam: John Benjamins.

Bryson, J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*. doi:10.1007/s10676-018-9448-6

Bryson, J. Diamantis, M. and Grant, T. (2017). Of, for and by the People: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25(3): 273-291

Carter, A. and Palermos, O. (2016). Is Having Your Computer Compromised a Personal Assault? The Ethics of Extended Cognition *Journal of the American Philosophical Association* 2(4): 542-560.

Chalmers, D. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Coeckelbergh, Mark. 2012. *Growing Moral Relations: Critique of Moral Status Ascription*. New York: Palgrave MacMillan.

Coeckelbergh, M. and Gunkel, D. (2014). Facing Animals: A Relational, Other-Oriented Approach to Moral Standing. *Journal of Agricultural and Environmental Ethics* **27(5): 715-733**

Coeckelbergh, M. and Gunkel, D. (2016). Response to ‘‘The Problem of the Question About Animal Ethics’’ by Michal Piekarski.. *Journal of Agricultural and Environmental Ethics* 29(4): 717-721

Danaher, J. (2018). Why we should create artificial offspring: Meaning and the Collective Afterlife. *Science and Engineering Ethics* 24(4): 1097-1118

Graham, G. (2015). Behaviorism. *Stanford Encyclopedia of the Philosophy*, available at https://plato.stanford.edu/entries/behaviorism/ (accessed 10/7/2018)

Gruen, L. (2017). The Moral Status of Animals. In Zalta (ed) *Stanford Encyclopedia of Philosophy* available at https://plato.stanford.edu/entries/moral-animal/

Guerrero, A. (2007). Don't know, don't kill: moral ignorance, culpability, and caution. *Philosophical Studies*, 136, 59–97

Gunkel, D. (2011). *The Machine Question*. Cambridge, MA: MIT Press.

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

Gunkel, D. (2018b). *Robot Rights*. Cambridge, MA: MIT Press.

Hauskeller, M. (2017). Automatic Sweethearts for Transhumanists. In Danaher, J. and McArthur, N. (eds) *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.

Hare, S. and Vincent, N. (2016) Happiness, Cerebroscopes and Incorrigibility: The Prospects for Neuroeudaimonia. *Neuroethics* 9 (1):69-84.

Holland, A. (2016). The Case Against the Case for Procreative Beneficence. *Bioethics* 30(7): 490-499

Jaworska, A. and Tannenbaum, J. (2018). The Grounds of Moral Status. In Zalta (ed) *Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/entries/grounds-moral-status/

Kaczor, C. (2011). *The Ethics of Abortion*. London: Routledge.

Leong, B. and Selinger, E. (2019). Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism" *2019 Proceedings of the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency*, pp 299-308.

Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingis, Trans.). Pittsburgh, PA: Duquesne University.

Levy, David. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics* 1 (3): 209–216. doi:10.1007/s12369-009-0022-6.

Lockhart, T. (2000). *Moral uncertainty and its consequences*. Oxford: OUP.

Moller, D. (2011). Abortion and moral risk. *Philosophy*, 86, 425–443

Neely, Erica L. (2014). Machines and the moral community. *Philosophy & Technology* 27 (1): 97–111. doi:10.1007/s13347-013-0114-y

Nyholm, S. and Frank, L.E. (2017). From Sex Robots to Love Robots: Is mutual love with a robot possible? In Danaher, J. and McArthur, N. (eds) *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.

Overall, C. (2011). *Why have children? The Ethical Debate*. Cambridge, MA: MIT Press.

Pardo, M. and Patterson, D. (2013). *Minds, Brains and Law*. Oxford, UK: Oxford University Press.

Puryear, S. (2017). Schopenhauer on the Rights of Animals. *European Journal of Philosophy* 25 (2):250-269

Raoult, A. and Yampolskiy, R. (2018). Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*, forthcoming – available at https://www.researchgate.net/publication/325498266_Reviewing_Tests_for_Machine_Consciousness (accessed 28/3/2019)

Regan, T. (1983) *The Case for Animal Rights*. University of California Press.

Saunders, B. (2015). Why Procreative Preferences May be Moral – And Why it May not Matter if They Aren't. *Bioethics* 29(7): 499-506.

Saunders, B. (2016). First, do no harm: Generalized Procreative Non-Maleficence. *Bioethics* 31: 552-558

Savulescu, J. (2001). Procreative Beneficence: Why We Should Select the Best Children. *Bioethics* 15: 413-426

Schwitzgebel, Eric, and Mara Garza. 2015. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy* 39 (1): 89–119. doi:10.1111/misp.12032.

Sebo, J. (2018). The Moral Problem of Other Minds. *The Harvard Review of Philosophy* 10.5840/harvardreview20185913

Singer, P. (1981). *The Expanding Circle*. Princeton, NJ: Princeton University Press.

Singer, P. (2009). Speciesism and Moral Status. *Metaphilosophy* 40 (3-4):567-58

Sparrow, Robert. 2012. Can machines be people? Reflections on the turing triage test. In *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. Patrick Lin, Keith Abney and George A. Bekey, 301–316. Cambridge, MA: MIT Press.

Stone, Z. (2017). Everything You Need To Know About Sophia, The World's First Robot Citizen. *Forbes* 7th November 2017, available at https://www.forbes.com/sites/zarastone/2017/11/07/everything-you-need-to-know-about-sophia-the-worlds-first-robot-citizen/#4e76f02b46fa (accessed 10/7/2018).

Sumner, L. (1987). *The Moral Foundations of Rights*. Oxford: Oxford University Press.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* 49: 433-460.

Tuvel, R. (2017). In Defence of Transracialism. *Hypatia* 32(2): 263-278

Vincent, J. (2017). Pretending to give robots citizenship helps no one. *The Verge* 30th October 2017, available at https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi-arabia-sophia (accessed 10/7/2018).

Warren, M.A. (2000). Moral Status: Obligations to Persons and Other Things. Oxford: Oxford University Press.

Weatherson, B. (2014). Running risks morally. *Philosophical Studies*, 167, 141–163