*Journal of Ethics and Social Philosophy*
VOL. 24, NO. 1 · FEBRUARY 2023
*https://doi.org/10.26556/jesp.v24i1.1659*


# CAN WE HAVE MORAL STATUS FOR ROBOTS ON THE CHEAP?

*Sebastian Köhler*

SHOULD ARTIFICIAL AGENTS (such as robots) be granted *moral status*? This seems like an important question to resolve, bearing in mind that we are very likely to encounter a growing number of increasingly sophisticated artificial agents in the not too distant future. Given that moral status is the property an entity has "if and only if *it* or *its interests* morally matter to some degree for the entity's own sake," without a clear answer about artificial agents' moral status, we risk either doing significant wrong in our interactions with such agents, or wasting significant moral concern that might be better allocated elsewhere.[1]

At the same time, many will think that before we can even start to tackle questions about the moral status of artificial agents, we first need to solve further, equally tricky issues in the philosophy of mind.[2] The "orthodox view" about moral status explains why:

> *Orthodoxy*: (Necessarily for all entities *e*) *e* has moral status only if it is (or belongs to a kind) capable of having mental states.[3]

If we accept Orthodoxy, we will think that for moral status considerations, what "goes on 'on the inside' matters greatly,"[4] as Nyholm and Frank put it. More precisely, we will believe that an entity such as an artificial agent will be eligible for moral status only if it has a mental life. As should be obvious, though, whether an entity has a mental life raises extremely controversial questions in

1 Jaworska and Tannenbaum, "The Grounds of Moral Status," emphasis added.

2 For an overview of the different prominent positions about moral status, see Jaworska and Tannenbaum, "The Grounds of Moral Status."

3 Note that the prominent views disagree about what kinds of mental states are relevant for moral status. What is relevant might be consciousness, the possession of preferences, or the possession higher level cognitive capacities. However, all of these views are committed at least minimally to Orthodoxy, which is why this paper focuses on this shared assumption.

4 Nyholm and Frank, "From Sex Robots to Love Robots," 223.

the philosophy of mind. Most importantly, this concerns the question of what it takes to have a mental life in the first place. Equally importantly, it also includes the question to what extent it is possible to have *evidence* that entities have a mental life. And this, to return to our specific context of artificial agents, raises the crucial question of whether what we *normally* take to be signs of a mind can also be taken as evidence for a mind across the board, including the *unnormal case* of artificial agents. As a result, if we are to follow Orthodoxy, there is no way around addressing such tricky questions in the philosophy of mind before settling the moral status of artificial agents, such as robots.

Given as much, one might hope that we could give a plausible account of moral status that evades these kinds of issues in the philosophy of mind that Orthodoxy imposes on us. Specifically, one might hope that there are plausible conditions on the basis of which we can grant entities moral status *without* having to undertake controversial commitments in the philosophy of mind. Let us call views of this kind "minimalism." Forms of minimalism are all views according to which we can give sufficient conditions on the basis of which we should grant entities moral status, without having to undertake commitments in the philosophy of mind. Note that minimalist views would have to satisfy at least two conditions to be plausible. First, they need to offer *sufficient* conditions for granting moral status to an entity that we can know to be satisfied without knowing whether that entity has a mind. Second, the plausibility of this criterion applied to concrete cases should not require us to combine the view with commitments in the philosophy of mind.

In recent years, a number of authors concerned with the moral status of artificial agents have suggested views that can, plausibly, be read in a minimalist vein.[5] This paper is concerned with the prospects of such views. However, the focus of the paper is not on minimalist views *generally*. Rather, the paper operates under a specific constraint. The paper brackets forms of minimalism about the moral status of artificial agents (such as those based on Luciano Floridi's *Information Ethics*, for example) that seem inherently overgeneralizing in the following sense: they imply that many more entities than we might think have moral status and, in particular, that many entities have moral status that are paradigmatic instances of entities lacking such status.[6] So, the paper focuses on views that are potentially the least revisionary forms of minimalism. Here, the paper looks at two recently suggested views that can plausibly be read

5 See, e.g., Danaher, "Welcoming Robots into the Moral Circle"; Coeckelbergh, "Growing Moral Relations"; Coeckelbergh and Gunkel, "Facing Animals"; Gunkel, "Robot Rights" and "The Other Question"; Floridi, "Information Ethics"; Tavani, "Can Social Robots Qualify for Moral Consideration?"

6 See Mosakas, "On the Moral Status of Social Robots," for arguments along those lines.

as views of this kind (and that have been getting some traction in the debate about the moral status of artificial agents), Ethical Behaviourism and Ethical Relationism.[7] These are here taken to be the *prima facie* most promising ways of fleshing out a minimalist view of this kind. This is so because—as will become more transparent in the discussion—they tie moral status to something it *is* paradigmatically related to, and, hence, are probably our best bet when it comes to developing forms of minimalism that are not radically revisionary (in what follows, I will use the label "minimalism" just to refer to minimalist views that are not revisionary in this sense).

With regards to these views, this paper argues that we should be pessimistic about the prospects of finding a plausible version of this kind of minimalism. Specifically, it argues that we have good reason to think that with regards to the two conditions on plausible minimalist views outlined above, views that satisfy the first condition will have to violate the second condition: views that satisfy the first condition are only plausible insofar as and because, in order to be applicable to concrete cases, they need to draw on additional assumptions in the philosophy of mind.[8] This is so, because (as I will try to show) our intuitions about moral status tend to *follow* our verdicts about what entities have a mind, and such views can accommodate this connection only by making assumptions about the mind. Hence, rather than avoiding controversies in the philosophy of mind, these views put themselves squarely within these controversies. As such, I conclude that there is no way of getting around the thorny questions in the philosophy of mind thought to plague Orthodoxy, when one is concerned with the moral status of artificial agents, unless one aims to *radically* revise the way we think about what entities have moral status.

The paper proceeds as follows: The first section discusses Ethical Behaviourism. The second Section discusses Ethical Relationism. The final section draws general lessons from the discussion and concludes.

## 1. ETHICAL BEHAVIOURISM

Ethical Behaviourism has recently been suggested by John Danaher.[9] He uses the position to argue that certain sorts of robots should be granted significant

7 For Ethical Behaviourism, see Danaher, "Welcoming Robots into the Moral Circle." For Ethical Relationism, see, e.g., Coeckelbergh, "Growing Moral Relations"; Coeckelbergh and Gunkel, "Facing Animals"; Gunkel, "Robot Rights" and "The Other Question."

8 Note that I am not arguing that these views or their proponents do *in fact* already make such assumptions. This is a matter of interpretation that I do not want to get into here. Rather, the point is that these views have to be combined with such assumptions to be plausible.

9 Danaher, "Welcoming Robots into the Moral Circle."

moral status. The view's core tenet is this: "If an entity *A* is *roughly performatively equivalent* to another entity *B* whom, it is widely agreed, has significant moral status, then it is right and proper to afford *A* the same moral status as *B*."[10] Danaher claims that Ethical Behaviourism is only an epistemic thesis: it is supposed to be a thesis about when we have *evidence* that an entity has moral status, not a thesis about when an entity has, in fact, moral status.[11] In fact, he claims that Ethical Behaviourism is, *in principle*, compatible with the thesis that moral status is *ultimately grounded* in having mental states.

It is important to note, however, that Danaher does not simply hold that performative equivalency is, e.g., *defeasible* evidence for an entity having moral status—which is something that a proponent of Orthodoxy could accept as well.[12] Rather, he holds a much stronger view:

> This article … 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 equivalency* between a robot and another entity to whom we afford moral status.[13]

That is, on Danaher's view performative equivalency is a *sufficient* condition for it to be the case that we should grant an entity a certain moral status: any entity that is performatively equivalent to another entity to which we grant moral status is to be granted the same moral status, no matter what other differences there might be between these entities. Hence, Ethical Behaviourism *starts* with a set of entities that are paradigmatic instances of entities with moral status—not considering in virtue of what they have moral status—and then assumes that we should expand the circle of which entities we attribute moral status to on the basis of performative equivalence, without considering any other differences there might be between these entities.

It should be clear that Ethical Behaviorism satisfies the first condition of minimalism, given that it provides us with a sufficient condition on the basis of

10 Danaher, "Welcoming Robots into the Moral Circle," 2025.

11 See Smids, "Danaher's Ethical Behaviourism," for issues with the epistemic reading of Ethical Behaviourism.

12 Note that, in fact, the view that performative equivalency is *defeasible* evidence for moral status is not a minimalist view at all. The best explanation as to why performative equivalency is defeasible evidence for moral status is that performative equivalency is defeasible evidence for mental equivalency. However, this position takes on distinctive commitments in the philosophy of mind, e.g., that what is evidence in normal cases for a mental life (e.g., other humans, non-human animals) is also such evidence in unnormal cases (such as artificial agents).

13 Danaher, "Welcoming Robots into the Moral Circle," 2025.

which we should grant moral status, and given that we can know whether this condition is satisfied without knowing whether the entities in question have a mind. After all, while all entities that are *paradigmatic* instances of entities with moral status are entities with a mind (namely humans and, maybe, some non-human animals), it is by no means clear that an entity that is performatively equivalent to an entity with a mind also has a mind. Furthermore, Danaher supports Ethical Behaviourism explicitly with the idea that it allows us to resolve tricky questions about whether we should grant moral status to robots without having to resolve tricky metaphysical and epistemological questions, e.g., in the philosophy of mind.[14] So, on the reading of Ethical Behaviorism suggested here, Ethical Behaviorism seems to aspire to being a form of minimalism.

At first sight, the criterion offered by Ethical Behaviourism is not implausible, at least in the sense that it is likely to give us intuitively correct verdicts with regards to moral status for a variety of normal cases. However, I will now argue that Ethical Behaviourism fails to offer a plausible version of minimalism, because the sufficient condition for granting moral status that it offers is only plausible to the extent that we combine it with certain assumptions in the philosophy of mind. If we drop these assumptions, the plausibility of Ethical Behaviourism disappears. Rather than avoiding controversial assumptions in the philosophy of mind, Ethical Behaviourism has to be combined with very robust such assumptions.

I will argue for this conclusion as follows: First, I will argue that on first sight Ethical Behaviourism faces a counterexample that is parallel to a counterexample to behaviourism in the philosophy of mind. I will then observe that Ethical Behaviourism escapes *this* counterexample only by making a move that is similar to a move made by those who adopt *functionalism* about the mind. This makes Ethical Behaviourism's verdicts about what entities have moral status co-extensional with functionalism's verdicts about what entities have a mind. I will then use a counter-example to *functionalism* about the mind to argue that this co-extensionality is *not* accidental, because the plausibility of Ethical Behaviourism depends on the plausibility of functionalism's verdicts. I will support this further by highlighting that if we adopt another view about the mind, teleo-functionalism, the plausibility of Ethical Behaviourism vanishes. Hence, to be plausible Ethical Behaviourism must be tied to functionalist assumptions in the philosophy of mind. While the argumentation here might seem more complicated than it needs to be to establish a problem for Ethical Behaviourism as a minimalist view, let me note that it takes this structure to draw out an

14 E.g., Danaher, "Welcoming Robots into the Moral Circle," 2028.

important further point for the discussion. This is that our intuitions about moral status tend to follow our verdicts about what entities have a mind.

As I said, my argumentation starts with a case from the philosophy of mind. This case is "Blockhead."[15] To start, consider the observation that "at any point in a creature's life there are only finitely many discriminately distinct possible inputs and outputs at its periphery."[16] If this is true, we could potentially draw a finite list of all the possible inputs and outputs in the life of any human. Let us call this list that human's "game of life." Plausibly, given any human's game of life, we could build a creature that is at first sight very similar to the human, but has a chip inside of it on which the human's game of life is inscribed, where this chip fully controls the behavior of the creature. Let us call such a creature the "Blockhead twin" of the human. Plausibly, any human has a possible Blockhead twin.

Blockhead twins are an objection to *behaviorism* in the philosophy of mind. Behaviorism in the philosophy of mind is the view that entities have mental states solely in virtue of certain behavioral dispositions. For example, on such a view an entity experiences pain just in case it displays the kind of *behavior* that is typically associated with a painful experience. The reason why Blockhead twins are relevant in the context of this paper is this: Blockhead twins *are* roughly performatively equivalent to human beings. Yet Blockhead twins clearly do not have a mind—behaviorism is false. And, they also should not be granted the moral status that a human being has. Assume, for example, that you could either save a human or their Blockhead twin from being crushed by a falling boulder. Here, it is intuitively very clear that you should save the human and not the Blockhead twin. In fact, Blockhead twins quite plausibly have *no* moral status—there is no question *at all* about whom to save in the described situation. And the best explanation for this is that Blockhead twins do not have a mind. Hence, Blockhead twins also seem to constitute a counterexample to Ethical Behaviourism.

At this point, Danaher has a response available, because on his view "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."[17] Hence, on his view Blockhead twins are not problematic, because while Blockhead twins' external physical behavior is performatively equivalent to that of humans, there is nothing in the twins that is performatively equivalent to the

15 See Block, "Psychologism and Behaviourism"; here I draw on the presentation in Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 114–19.

16 Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 116.

17 Danaher, "Welcoming Robots into the Moral Circle," 2028.

functional operations of the brain. And this explains why Blockhead twins should not be granted the moral status that humans have.

However, in making this move, Danaher makes the verdicts about what entities should be granted moral status co-extensional with what entities have mental states according to *functionalists* about the mind.[18] According to functionalists, mental states are characterized by their relational properties, namely by their relations to inputs to and outputs from the cognitive system (what would matter on a purely behaviorist picture) and their relations amongst each other. That is, "a functionalist theory of mind specifies mental states in terms of three kinds of clauses: input clauses that say which conditions typically give rise to which mental states; output clauses that say which mental states typically give rise to which behavioural responses; and interaction clauses which say how mental states typically interact."[19] Hence, for a functionalist Blockhead twins do not have a mind, because they lack the relevant *internal* functional architecture, while anything that has the relevant functional architecture and inputs and outputs does have a mind. So, anything that should be granted moral status on the sophisticated version of Ethical Behaviourism is something that has a mind according to functionalism.

Co-extensionality by itself is not a problem. However, I will now argue that it is not a coincidence, because the plausibility of Ethical Behaviourism *depends* on the plausibility of functionalism's verdicts. To illustrate this, I will use another case from the philosophy of mind, the Global Brain.[20] Assume that there is a computer program that mimics the functional operation of a human brain at neuron by neuron level. This program is run as follows: First, each adult living on Earth is assigned the job of just one neuron (this is a highly unrealistic assumption, as even the Earth's population is not large enough). Each earthling gets a phone for this specific purpose and what they have to do is to call certain numbers, if they are called by other numbers. To tell them what to do, they have a specific list of instructions that "exactly models what their assigned neuron does, and the inputs to and outputs from their phones are connected up so as to run the program."[21] Furthermore, the network is connected to a robot body, which receives inputs from its environment in a similar way these inputs come

18 For an introduction and overview, see Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 45–64, 84–94, and 107–28; or Levin, "Functionalism."

19 Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 47.

20 The example originates in Block, "Troubles with Functionalism." Here I draw on the updated version in Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 107-8. Note that the original example is called the "China Brain." I've modified it to avoid problematic stereotyping or exoticizing that might be present in that version.

21 Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 107.

to us. The network also produces outputs that go from the robot body that, then, in actual and counterfactual circumstances behaves very similarly to a human. "Thus, the android will behave in the various situations that confront it very much as we do, despite the fact that the processing of the environmental inputs into final behavioural outputs goes via a highly organized set of [earthlings] rather than a brain."[22]

Let us call the system that consists in the robot plus the population of Earth the "Global Brain." The Global Brain is a counter-example to functionalism, because functionalism implies that Global Brain has the same mental states as a human, while it seems intuitively that it does not have mental states at all. Assume for the moment with functionalism's opponents that it is, in fact, intuitively plausible that Global Brain does not have mental states. Then it *also* seems plausible that it does not have moral status: Global Brain should *not* be granted the same moral status that is granted to a human. For example, if we face a choice between killing a human or shutting down Global Brain, then it is intuitively very clear that you should shut down Global Brain. In fact, Global Brain quite plausibly has *no* moral status, as there is no question *at all* about whom to save in this situation. Hence, the Global Brain *also* constitutes a counterexample to Ethical Behaviourism. But why does it not seem plausible that Global Brain has moral status? As with the Blockhead twin, the best explanation for this seems to be that the Global Brain does not have a mind.

Here is a way to substantiate that this is the best explanation: functionalists will deny the intuitions about Global Brain. They will try to argue that if we only think about the case in the right way, we will think that Global Brain *does* have the same mental states as a human.

> The source of the intuition that [Global Brain] lacks mental states like ours seems to be the fact that it would be so *very* much bigger than we are. We cannot imagine "seeing" it as a cohesive parcel of matter. . . . A highly intelligent microbe-sized being moving through our flesh and blood brains might have the same problem. It would see a whole mass of widely spaced entities interacting with each other in a way that made no sense to it.[23]

But notice what happens if we submit to this resistance by functionalists, and come to believe that Global Brain has a mind just like any human does. In this case, our verdicts about moral status will tend to align with the verdicts of Ethical Behaviourism: we will likely think that Global Brain *should* be afforded the

22 Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 108.

23 Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 109.

same moral status that is afforded to humans, making the case described above a real moral dilemma. So, even if our intuitions about Global Brain having a mind are, in the end, incorrect, this case and the Blockhead case still spell trouble for Ethical Behaviourism as a form of minimalism: What these cases reveal is that the plausibility of Ethical Behaviourism seems to hang on the commitment to some sort view in the philosophy of mind, namely one that entails that the mental lives of two *roughly performative equivalent* beings are identical. Specifically, it seems as if Ethical Behaviourism stands or falls with the plausibility of such a view about the mind. So, Ethical Behaviorism seems to violate the second condition for a plausible version of minimalism.

We can bring this point home by considering views about the mind which do not entail that the mental lives of two *roughly performative equivalent* beings are identical. Consider, for example, teleo-functionalist views.[24] According to these views, in biological beings like us etiological biological functions are (at least partially) constitutive of having mental states, where the etiological function of something is what explains why it exists. On such views, we have mental states *because* of our biological history—*because* us having such states was selected by evolution. But note that on such a view, a being with a different causal history from ours could lack mental states, even if it is *roughly performative equivalent* to us. So, for example, if such a view was true, even a highly sophisticated android that is by all plausible performative standards roughly equivalent to a human—such as Data from *Star Trek*—would not have *any* mental states.[25]

Suppose now that we believe teleo-functionalism to be true, and so concede that Data does not have mental states. In this case, we would likely also think that the robot should not be granted the same moral status that we possess.[26] After all, despite all appearances, this robot would not experience pain, have preferences or interests, or anything else for which one needs a mental life. And in this case, it seems intuitively quite clear that none of the moral reasons

24 For an introduction and overview of teleo-functionalist views, see Neander and Schulte, "Teleological Theories of Mental Content."

25 See Hofmann, "Could Robots Be Phenomenally Conscious."

26 Danaher considers the objection that the efficient cause of an entity might matter (Danaher, "Welcoming Robots to the Moral Circle," 2032-3). Specifically, he considers whether it might be relevant that humans (or animals) are produced by evolution, while robots are manufactured. He rejects this objection by pointing out that humans or animals could also be manufactured. However, this response would miss the central point of the objection pressed here, which is that an entity with a different kind of causal history lacks mental states. A human that is the product of selective breeding still stands in the relevant evolutionary chain to make it such that she has mental states, but a robot that is manufactured stands in no such chain.

that derive from moral status are present. However, what this means is that if a teleo-functional view about the mind is correct, Ethical Behaviourism must be false. And that means that to be plausible, Ethical Behaviourism must be combined with a commitment to the idea that teleo-functionalism is false, and something like the functionalism presented further above is true. Hence, rather than avoiding controversy in the philosophy of mind, Ethical Behaviourism is hostage to fortune with regards to such controversies.

What all of this shows is this: Ethical Behaviourism is plausible as a thesis about what entities should be granted moral status only to the extent that we combine it with certain assumptions in the philosophy of mind. That is, Ethical Behaviourism is only plausible to the extent that what it identifies as a sufficient condition on the basis of which we should grant moral status is also a sufficient condition for assuming that an entity having a mind. So, rather than it not mattering what goes on "on the inside," this actually matters a great deal for the plausibility of Ethical Behaviourism. It turns out, therefore, that Ethical Behaviourism is only really plausible if it is paired with very robust assumptions in the philosophy of mind. This means that Ethical Behaviourism does not really allow us to avoid the tricky questions in the philosophy of mind that plague Orthodoxy. At most, it assumes these questions away. So, Ethical Behaviourism cannot offer a plausible form of minimalism.

Furthermore, there is a more general lesson to take away from this discussion, namely this: as cases like Blockhead twins and Global Brain illustrate, our verdicts about what entities should or should not be afforded moral status, and about what entities have a mind tend to go hand in hand. When we judge an entity to not have a mind, we will tend to judge that it does not have moral status. A plausible view about moral status needs to be able to accommodate this. It is unclear how forms of minimalism can accommodate this, however, unless they are combined with assumptions about the mind that make their verdicts about moral status coincide in the relevant cases with what entities have a mind. If this is the case, though, these positions would not actually offer progress over Orthodoxy, at least when it comes to evading tricky issues in the philosophy of mind. But Ethical Behaviourism is not the only theory to which this applies. It holds for Ethical Relationism too, as I will argue next.

## 2. ETHICAL RELATIONISM

In recent work, Mark Coeckelberg and David Gunkel have suggested what they call a "relational, other-oriented approach to moral standing."[27] The position

27 This is, in fact, the subtitle of Coeckelbergh and Gunkel, "Facing Animals."

they develop (in co-authored and individual work) is very rich and complex, and it will not be possible to do justice to all of the details here.[28] What I will do is present what I take to be a plausible version of the view, focusing on the details that are central for the discussion here.

The core idea of Ethical Relationism is this:

> As we encounter and interact with others—whether they be other human persons, an animal, the natural environment, or a social robot—this other entity is first and foremost situated in relationship to us. Consequently, the question of social and moral status does not necessarily depend on what the other is in its essence but on how she/he/it … supervenes before us and how we decide, in "the face of the other" (to use Levinasian terminology), to respond.[29]

How should we understand this? First, the basic idea is that entities have moral status in virtue of the *relations* in which we stand to them. What kinds of relations? The general idea seems to be the following: in our interactions with certain other entities, these entities *present* themselves to us in certain ways. When such an entity presents itself to us as "having face," it has moral status.

This terminology of something "presenting itself to us" or "having face," needs a little unpacking. As I understand it, these notions are supposed to be graspable by the distinctive *phenomenology* of coming to regard an entity as worthy of moral consideration. This is the phenomenology of, e.g., experiencing discomfort at the entity being treated in certain ways, feeling affection for the entity, regarding the entity with respect, etc. that might arise in us due to our interactions with an entity.

To make the phenomenology vivid, think about the Robovie experiment that tested how children relate to social robots.[30] In this experiment, children ages nine to fifteen interacted with a humanoid robot for roughly fifteen minutes, including making small talk with the robot, playing a game, and the robot asking the child for a hug. At end of the session an adult comes and orders the robot to go into a closet. The robot protests vehemently, asking not to be put in the closet, which it claims is dark and lonely. Now put yourself in the position of a child having interacted with Robovie in this way. It seems plausible to assume that you will feel a certain bond with Robovie, and that you will feel distress at Robovie's reaction to being put in the closet. And, if you

28 They have developed and defended the view in various places, e.g., Coeckelbergh, "Growing Moral Relations"; Coeckelbergh and Gunkel, "Facing Animals"; Gunkel, "Robot Rights." The view draws on the work of Emmanuel Levinas. See Levinas, "Totality and Infinity."

29 Gunkel, "Robot Rights," 96.

30 Kahn et al., "Robovie, You'll Have to Go into the Closet Now."

watch videos of the experiment you will likely find this reaction compelling—an emotional reaction that lingers on, even when you know that Robovie is nothing more than a puppet remotely controlled by a human in another room.[31] As I understand Coeckelbergh and Gunkel, these sorts of phenomenological responses are what determine moral status. Specifically, it is when these phenomenological responses make us experience the entity as "having face" that it should be afforded moral status—where I assume the phenomenology of experiencing something as "having face" is what we *paradigmatically* experience when we come to regard something as worthy of moral consideration. Hence, for Coeckelbergh and Gunkel the phenomenological responses of coming to regard something to be worthy of moral consideration in our interactions with it have *explanatory priority* to being of moral consideration. This is a reversal of the ordinary order of explanation: normally we would think that something being of moral consideration is what the relevant sorts of responses that constitute moral regard are *sensitive* to, while here being of moral consideration is determined or constructed by these responses.

Cockelbergh's and Gunkel's own presentation of the account raises many questions that I do not have room to investigate here. The way I read it, though, a helpful analogy for understanding this sort of view can be found in secondary quality accounts of colors.[32] On such accounts, something having a certain color crucially depends on our own responses to the thing. For example, on such accounts, things are red in virtue of appearing red to certain sorts of observers in certain sorts of conditions. Similarly, according to Ethical Relationism an entity has moral status in virtue of triggering the kinds of responses that constitute moral regard in certain sorts of observers in certain sorts of conditions. This might not be Coeckelbergh's and Gunkel's official account.[33] However, this is the reading I will presuppose in what follows to sharpen the discussion.

Ethical Relationism seems to be a form of minimalism. After all, we can *in principle* come to experience the distinctive phenomenology of moral regard toward an entity in our interactions with it without knowing whether it has a mind. The Robovie case illustrates this clearly: Robovie does not have a mind, but it presents itself in a way that gives rise to the right kinds of phenomenological responses. So, Robovie could, *in principle*, present as having moral status,

31 For illustration, see, e.g., http://depts.washington.edu/hints/video8.shtml.

32 For a secondary quality account of color, see, e.g., Johnston, "How to Speak of the Colors." Another helpful analogy might be the Strawsonian account of responsibility (see Strawson, "Freedom and Resentment") on which being responsible just *is* being the proper target for our reactive attitudes.

33 Coeckelbergh alludes to this analogy (see "Growing Moral Relations," 207), though it seems that he does not fully embrace it.

even though it has no mind. And, Coeckelbergh and Gunkel explicitly argue that a primary motivation for Ethical Relationism is that it avoids the tricky questions in the philosophy of mind that plague Orthodoxy.[34] However, using the lessons from the discussion of Ethical Behaviourism, I will now argue that Ethical Relationism also does not provide a plausible version of minimalism, because Ethical Relationism is only plausible if it is combined with additional assumptions in the philosophy of mind. Again, therefore, the second condition that plausible versions of minimalism need to satisfy is violated.

The way to argue for this is, simply, by alluding to the cases we've discussed before, and by asking what best explains what is going on in these cases—which is something any plausible view about moral status should capture. Think about Blockhead twins or the Global Brain. What the discussion of Ethical Behaviorism revealed is that, in these cases, our intuitions about whether the entities in question should be afforded moral status follow our verdicts about whether they have a mind. If we meet an entity like Blockhead Twin or Global Brain, we might at first be inclined to attribute moral status to it, but we will retract this verdict upon realizing that the entity, plausibly, does not have mental states. In fact, it seems that Robovie also supports this. Our initial responses to Robovie upon interacting with it are, plausibly, the kinds of responses that characterize regarding a being as worthy of moral consideration. But, when someone shows us the remote control and explains that Robovie is, actually, just a puppet, we will judge that Robovie is not actually worthy of moral consideration.

How can Ethical Relationists accommodate these cases, and the way our intuitions about moral status here follow our verdicts about who has a mind? First, what the cases show is that the kinds of responses that characterize seeing a being as worthy of moral consideration *by themselves* cannot be sufficient for it to be the case that we *should* afford the being moral consideration. After all, it would be implausible to hold that Blockhead Twins, Global Brains, or Robovie do have moral status as long as we are mistaken about their mental lives, and they lose their moral status when we come to think that nothing is going on "on the inside."

However, Ethical Relationism as I understand it here already accommodates this. After all, on the suggested version of Ethical Relationism, an entity has moral status in virtue of triggering moral regard in *certain sorts of observers, in certain sorts of conditions*. So, with the right kind of view about what sorts of observers and what sorts of conditions are relevant here, Ethical Relationism might be able to explain how and why our verdicts about moral status and mental life tend to go together, and to imply the correct verdicts about the moral status of entities such as Blockhead twins, Global Brain, and Robovie.

34 See, e.g., Coeckelbergh and Gunkel, "Facing Animals," 718–20.

What sorts of conditions and observers might an Ethical Relationist invoke at this point, though? Here, I will offer what seems to me the most plausible and straightforward response to this question.

Let me start with an observation about the responses that characterize moral regard. Irrespective of what we think about the relationship between these responses and something having moral status, it is, *prima facie*, very plausible that there is a close connection between many of the most central kinds of responses that constitute moral regard and our inclinations to attribute mental states. A core part of moral regard is to *feel with* and *feel for* the entity in question, to put ourselves in that entities' shoes, and so on. Hence, there seems to be a tight connection between our mindreading capacities and some of the most central responses that constitute moral regard. Specifically, it seems that many of the core responses that constitute moral regard are partially constituted by or are the result of our mindreading capacities.

However, our mindreading capacities sometimes *misfire*. For example, it is likely that this will happen when we interact with Blockhead twins. In fact, this is what happens with Robovie. And it is unsurprising that our mindreading capacities misfire in these sorts of situations. After all, while our mindreading capacities probably work very well in normal conditions where the relevant sorts of moral regard are good evidence for the presence of minds, the relevant sorts of cases are *not* normal: they are not situations for which our epistemic capacities for gaining knowledge of other minds have evolved.

These observations allow us to suggest a diagnosis as to what is going on in the cases in question: plausibly, the kind of moral regard we experience in reaction to Blockhead twins, Global brain, or Robovie is exactly the kind triggered or partially constituted by our mindreading capacities. But, when we realize that these relevant capacities *misfired* in this case, we retract our inclination to attribute moral status. If this explanation is plausible (and I cannot see a better explanation), this, in turn, allows us to spell out further the *conditions* in which something has to appear to certain observers as having moral status in a way that allows us to account for the intuitions in question. This suggests that part of the relevant conditions for granting moral status is just this: that our epistemic capacities for gaining knowledge of other minds do *not* misfire.

If we include this sort of condition in Ethical Relationism, we can explain why our intuitions about who has a mind and our intuitions about who should be afforded moral status tend to go together: those responses of moral regard that result from our mindreading capacities are amongst our most central ones, and their sensitivity to these capacities explains the connection. We can offer this explanation only, however, by making assumptions about when and why our mindreading capacities misfire—i.e., by making assumptions about when

entities have a mind and when we have knowledge of their minds. Hence, in including this sort of condition, we've now uncovered that to deal with the challenge on the table, Ethical Relationism has to violate the second condition for a plausible form of minimalism: to make plausible verdicts about certain sorts of cases, Ethical Relationism must be combined with certain robust assumptions in the philosophy of mind. But, this puts the account squarely back into thorny philosophical issues in the philosophy of mind regarding these capacities. Hence, rather than avoiding such problems, we are back to facing these problems.

At the end of section one it was observed that a plausible view about when we should grant entities moral status needs to be able to accommodate the fact that our intuitions about moral status and about who has a mind tend go together. It was suggested that it is unclear how forms of minimalism can accommodate this unless they make assumptions about the mind. The discussion of Ethical Relationism in this section substantiates this suspicion. Before I go on to draw some general lessons from the discussion, though, let me respond to two possible worries about my argument against Ethical Relationism.

First, we should address whether the argumentation against Ethical Relationism is question-begging. Does the argument here implicitly *presuppose* that only entities with mental states have moral status?[35] Two features of my argumentation might raise this worry. First, above I have tied our mindreading capacities to moral regard. The way I did this might give the impression that the argumentation already presupposes that *proper* moral regard is only triggered by entities with a mind. Second, the discussion did not even consider any "mind-less" entities that might conceivably have moral status, such as plants, forests, mountains, or the planet itself, that explicitly figure in some of Coeckelbergh's and Gunkel's work.[36] This might give the impression that the cases I've focused on smuggle in, again, the assumption that only entities with mental states have moral status into the argumentation.

However, my argumentation does not rely on such an illicit implicit assumption. First, the challenge that I raised for Ethical Relationism does not itself rest on such an assumption. What the challenge does is ask Ethical Relationism to explain why our intuitions about what entities have a mind and what entities have moral status tend to go together in the way suggested by cases like Blockhead twins, Global brain, or Robovie. This challenge would stand, even if there *were* clear cases in which our verdicts about moral status are not influenced by our judgements about minds. So, the challenge itself

35 Thanks to an anonymous reviewer for *JESP* for suggesting that I address this worry.

36 See, e.g., Coeckelbergh, "What Do We Mean by a Relational Ethics?" and "Environmental Skill"; Gunkel, "Robot Rights."

is compatible with the possibility that mindless things have moral status: to raise the challenge, we do not have to implicitly assume that only entities with mental states have moral status.

Second, the response to the challenge that I offered on behalf of Ethical Relationists is also not tied to a question-begging assumption. I suggested that Ethical Relationists can deal with the challenge by imposing a condition on the kinds of responses that matter for determining moral status: that the reactions that matter are formed by observers in conditions in which our mindreading capacities do not misfire. However, this condition is *compatible* with people reacting with moral regard to at least certain things that do not have a mind. Whether there are such cases depends on what kinds of moral regard there are, and what they are triggered by. Regarding this, the explanation only takes the following stance: First, there is a close connection between many responses that constitute moral regard and our inclinations to attribute mental states. Second, these responses are *core* to what constitutes moral regard. Third, focus on these responses is sufficient to explain what is going on in the cases relevant for the discussed challenge. All of these claims are very plausible, but none of them rules out the possibility of other kinds of moral regard triggered by mindless entities. Specifically, the explanation does not need the assumption that *all* types of moral regard are tied to our attributions of mental states. It only needs the assumption that this is true for *some* types, and that appeal to those is sufficient to deal with the challenge. The explanation is compatible, for example, with there being a distinctive type of *moral* regard we experience in relation with a magnificent oak. Note, though, that even if there *are* such responses, the condition I suggested for dealing with the challenge is not problematic. These responses are not going to be influenced by our mindreading capacities either way, so requiring that moral status be determined by the responses of people in conditions in which these capacities do not misfire is not going to change what moral status is attributed on the basis of these responses. Hence, the argument against Ethical Relationism does not rely on question-begging assumptions about the connection between minds and moral regard.

Before I move on to the second worry, let me end the discussion here by flagging that even if there *are* types of moral regard properly triggered by mindless entities, they are not going to help *minimalist* Ethical Relationism. First, these responses do not change the fact that the best way for Ethical Relationists to deal with the challenge is to add the condition I suggested. But, adding this condition to Ethical Relationism itself defeats the minimalist aspirations of the view. Second, even if we *only* focus on the case of artificial agents, the relevant types of responses are not going to be of help. This is so because the kind of moral regard that is triggered by such agents is, very plausibly, exactly the type

of moral regard that is partially constituted by, or the result of, our mindreading capacities. This is suggested, for example, by reflection on the cases discussed in this paper, and on what best explains how our intuitions about these cases are influenced by our verdicts about the presence of minds.

Let me now discuss a second potential worry. This is the worry about whether I've discussed the best version of Ethical Relationism here.[37] The view I've discussed draws on work by Cockelbergh and Gunkel, but their actual view is more complex than the one I suggest. Most relevantly, Coeckelbergh's and Gunkel's work highlights certain sorts of factors that influence moral status attributions, but which have not been considered as part of the account I suggest. The main factors they focus on are the language we use to talk about entities, our relationships with those entities, and social norms concerning those entities. Coeckelbergh and Gunkel note, for example, that it matters a lot for how one feels about an animal whether one tends to think of it "as an animal," or whether one has given it a name. Similarly, social acceptance of eating meat can influence how we feel for farm animals in the way we engage with them. This raises the question of whether these additional factors can help Ethical Relationists with the challenge at hand.

No. First, let us be clear how exactly these observations should feed into the account. Coeckelbergh's and Gunkel's observations are about our actual responses of moral regard, and such factors do indeed, very plausibly, influence our actual responses of moral regard in our interactions with entities. However, what should matter on the Ethical Relationist account are not our *actual* responses of moral regard in response to our interactions with an entity: this would be very implausibly relativistic *and* highly revisionary.[38] It would imply, implausibly, for example, that if slave owners harden themselves to the plight of the enslaved, slaves do not have moral status, at least relative to slave owners. Rather, it must be the responses of people who interact with the entity under certain *ideal* conditions that determine moral status. I assume that this should mean at least that one reacts to the entity as a result of *fair* engagement:

37 Thanks to an anonymous reviewer for *JESP* for raising this objection.

38 In fact, it is important to be very clear here that the only thing that matters is the influence these factors have on *moral* regard, not on other emotional reactions. After all, these factors influence a variety of ways we feel about entities. For example, the relationship I have with my wedding ring makes it such that I care a lot about it—much more than I care about qualitatively identical other rings. However, the way my relationship makes me feel about my wedding ring should not count for moral status. This would be *highly* revisionary: my wedding ring matters *to me*, but it does not matter for its own sake. Of course, this raises another issue for Ethical Relationism that I have not talked about, which is how we individuate specifically *moral* regard. But this is an issue that goes beyond this paper. Here we should rely on the intuitive difference between different types of responses.

engagement that is honest and freed from certain pre- and mis-conceptions that unduly interfere with the responses that constitute moral regard. I think that this sort of emphasis on fair engagement is a plausible reading of how Coeckelbergh and Gunkel themselves would understand their suggestions.

The factors they highlight as influencing moral status attributions can then be understood as feeding into the conditions in which moral regard has to be formed to determine moral status. This is plausible: if our naming practices interfere with the moral regard we feel for animals, for example, then only responses should count that are not unduly influenced by such practices. Of course, figuring out how exactly to spell out undue interference by these sorts of factors in a way that is non-circular is going to be quite a challenge for Ethical Relationists, but this is not an issue that should concern us here.

What is important here is only that we can now respond the second worry raised above, as we can now see that these additional factors are not going to help with the challenge. For cases like Blockhead Twins, Global Brain, or Robovie, it does not seem plausible that social norms, relationships, or the language we use interfere unduly with our moral regard. In fact, if there is *anything* that interferes in these cases with our responses, it is our overreactive tendency to attribute minds to mind-less entities. A good example for this, I think, is the case of the social humanoid robot Sophia.[39] Sophia *does* have a name and if we consider the way people tend to interact with it, it seems quite plausible that Sophia triggers in them at least some of the responses we associate with moral regard. However, when we understand the actual workings of the robot and that it is little more than a "chat-bot with a face," our inclination to attribute moral status immediately disappears.[40] It would be implausible, though, to attribute this retraction to an interference of the language we use or the relationships we have with the robot. Rather, what explains the retraction best, quite simply, is that our earlier attribution was based on our now corrected misconception that Sophia has a mental life.[41] One might suggest that we are unduly influenced here by a preconception that Sophia would need to have a mind to have moral status, but now *this* would be question-begging without further argument that it is an *illicit* preconception.

39 See Hanson Robotics, "Sophia."

40 See, e.g,. Gershgorn, "Inside the Mechanical Brain of the World's First Robot Citizen"; Ghosh, "Facebook's AI Boss Described Sophia the Robot as 'Complete B------t.'"

41 This impression is further suggested by, e.g., the way David Hanson (the founder of Hanson Robotics which created Sophia) tends to (mis-)represent Sophia in a way that suggests the existence of a mental life (see Vincent, "Sophia the Robot's Co-Creator Says the Bot May Not Be True AI"; Hanson Robotics marketing material (see Hanson Robotics, "Sophia") is also very suggestive in this regard).

In any case, it does not seem as if the further conditions that we might derive from Coeckelbergh and Gunkel would help Ethical Relationism deal with the challenge that I posed. Rather, it still seems like the best way to accommodate the way our intuitions about moral status tend to follow our verdicts about who has a mind is to build the condition I suggested into Ethical Relationism. But this condition saddles Ethical Relationism with commitments in the philosophy of mind.

## 3. Conclusion: General Lessons about Minimalism

The discussion of this paper has yielded two important lessons. First, that our intuitions about what entities should be afforded moral status tend to go hand in hand with our intuitions about whether an entity has a mind. In a sense, this is unsurprising, given how strong Orthodoxy's following is. It is a significant result in the face of the promises of minimalism, however, because with this result on the table, one can formulate a challenge to any form of minimalism: it needs to be able to accommodate how our intuitions go together in this way, and to explain why those intuitions tend to go together.

The second important lesson is then about the prospects of minimalism meeting this challenge. Here the discussion provides evidence that views about moral status that do not explicitly subscribe to a version of Orthodoxy can accommodate these intuitions about moral status only by themselves being combined with controversial assumptions in the philosophy of mind.[42] And this means that such views violate the second condition for plausible versions of minimalism. Rather than avoiding controversies in the philosophy of mind, such views put themselves squarely within those controversies.

In this paper we have only looked at views that try to give sufficient conditions for granting moral status in terms of two sorts of alternatives to mental states: outward behavior and moral responses. While this set of alternatives is restricted, these are also the *prima facie* most promising alternatives, as they tie

42 I am proceeding on the assumption, of course, that we take the intuitions to be legitimate evidence for moral status. Another option for the minimalist to respond to the challenge is to debunk those intuitions, e.g., by suggesting they are based in some sort of bias. Obviously, though, minimalists would have to give us good reasons to think that these intuitions need to be debunked—our moral intuitions are amongst our best guides to the correct answers to moral questions and we should not just give them up just because they conflict with minimalism. I don't see any good reasons to give these particular intuitions up (in particular because there are many ways to explain why there might be moral requirements to interact in certain ways with entities without a mind, even if we concede they lack moral status. Things that do not matter in themselves might still matter because people care about them, after all).

moral status to something it *is* paradigmatically related to: both are *normally* very good evidence for moral status.[43] If even these kinds of views do not succeed—which are probably our best bet when it comes to developing forms of minimalism that are not radically revisionary—this is strong evidence that when thinking about moral status, there is no way around the issues in the philosophy of mind that plague Orthodoxy (at least not without radical revision). Hence, these sorts of issues do not seem to provide a good motivation to move away from Orthodoxy.[44]

*Frankfurt School of Finance and Management*
*s.koehler@fs.de*

## REFERENCES

Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.

Block, Ned. "Psychologism and Behaviourism." *The Philosophical Review* 90, no. 1 (January 1981): 5–43.

Braddon-Mitchell, David, and Frank Jackson. *The Philosophy of Mind and Cognition*. 2nd ed. Oxford: Blackwell, 2007.

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

Coeckelbergh, Mark. *Environmental Skill. Motivation, Knowledge and the Possibility of a Non-Romantic Environmental Ethics*. New York: Routledge, 2015.

Coeckelbergh, Mark. "What Do We Mean by a Relational Ethics? Growing a Relational Approach to the Moral Standing of Plants, Robots and Other Non-Humans." In *Plant Ethics. Concepts and Applications*, edited by Angela Kallhoff, Marcello Di Paola, and Maria Schörgenhumer, 98–109. New York: Routledge, 2018

Coeckelbergh, Mark, and David Gunkel. "Facing Animals: A Relational,

43 Of course, the discussion here indicates that the primary reason why these are *normally* good evidence for moral status just is because they are *normally* good evidence for a mental life. The problem with artificial agents, however, is that it is not clear whether what *normally* is good evidence for a mental life is also evidence for a mental life in conditions that are not normal.

44 For helpful feedback I would like to thank Johannes Himmelreich, Leo Menges, Christine Tiefensee, as well as two anonymous referees for this journal. I would also like to thank the audience of the FS Philosophy Forum, where the idea for this paper was born. Research for this paper was conducted while I was a principal investigator of the project group *Regulatory Theories of AI* of the *Centre Responsible Digitality* (ZEVEDI).

Other-Oriented Approach to Moral Standing." *Journal of Agricultural and Environmental Ethics* 27, no. 5 (October 2014): 715–33.

Danaher, John. "Welcoming Robots into the Moral Circle: A Defence of Ethical Behavourism." *Science and Engineering Ethics* 26, no. 4 (August 2020): 2023–49.

Floridi, Luciano. "Information Ethics: On the Philosophical Foundation of Computer Ethics." *Ethics and Information Technology* 1, no 1 (March 1999): 33–52.

Gershgorn, Dave. "Inside the Mechanical Brain of the World's First Robot Citizen." Quartz. November 12, 2017. https://qz.com/1121547/how-smart-is-the-first-robot-citizen/.

Ghosh, Shona. "Facebook's AI Boss Described Sophia the Robot as 'Complete B------t' and 'Wizard-of-Oz AI.'" Business Insider. January 6, 2018. http://www.businessinsider.com/facebook-ai-yann-lecun-sophia-robot-bullshit-2018-1.

Gunkel, David. *Robot Rights*. Cambridge MA: MIT Press, 2018

Gunkel, David. "The Other Question: Can and Should Robots Have Rights?" *Ethics and Information Technology* 20, no. 2 (June 2018): 87–99.

Hanson Robotics. "Sophia." https://www.hansonrobotics.com/sophia/.

Hofmann, Frank. "Could Robots Be Phenomenally Conscious." *Phenomenology and the Cognitive Sciences* 17, no. 3 (July 2017): 579–90.

Jaworska, Agnieszka, and Julie Tannenbaum. "The Grounds of Moral Status." *The Stanford Encyclopedia of Philosophy* (Spring 2021). https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/.

Johnston, Mark. "How to Speak of the Colors." *Philosophical Studies* 68, no. 3 (December 1992): 221–63.

Kahn, Peter H., Takayuki Kanda, Hiroshi Ishiguro, Nathan G. Freier, Rachel L. Severson, Brian T. Gill, Jolina H. Ruckert, and Solace Shen. "'Robovie, You'll Have to Go into the Closet Now': Children's Social and Moral Relationships With a Humanoid Robot." *Developmental Psychology* 48, no. 2 (March 2012): 303–14.

Levin, Janet. "Functionalism." *The Stanford Encyclopedia of Philosophy* (Winter 2021). https://plato.stanford.edu/archives/win2021/entries/functionalism/.

Levinas, Emmanuel. *Totality and Infinity: An Essay on Exteriority*. Pittsburgh: Duquesne University Press, 1969.

Mosakas, Kestutis. "On the Moral Status of Social Robots: Considering the Consciousness Criterion." *AI and Society* 36, no. 2 (June 2021): 429–43.

Neander, Karen and Peter Schulte. "Teleological Theories of Mental Content." *The Stanford Encyclopedia of Philosophy* (Spring 2021). https://plato.stanford.edu/archives/spr2021/entries/content-teleological/.

Nyholm, Sven, and L E. Frank. "From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?" In *Robot Sex: Social and Ethical Implications*, edited by John Danaher and Neil McArthur, 219-244. Cambridge MA: MIT Press, 2017.

Smids, Jilles. "Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot?" *Science and Engineering Ethics* 26, no. 5 (October 2020): 2849–66.

Strawson, P. F. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 1–25.

Tavani, Herman T. "Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights." *Information* 9, no. 4 (April 2018): 73.

Vincent, James. "Sophia the Robot's Co-Creator Says the Bot May Not Be True AI, but It Is a Work of Art." *The Verge*. (November 2017). https://www.theverge.com/2017/11/10/16617092/sophia-the-robot-citizen-ai-hanson-robotics-ben-goertzel.