AI and Ethics (2024) 4:213–228
https://doi.org/10.1007/s43681-023-00260-1

ORIGINAL RESEARCH

What would qualify an artificial intelligence for moral standing?

Ali Ladak[1,2]

Received: 6 November 2022 / Accepted: 5 January 2023 / Published online: 25 January 2023

Abstract

What criteria must an artificial intelligence (AI) satisfy to qualify for moral standing? My starting point is that sentient AIs should qualify for moral standing. But future AIs may have unusual combinations of cognitive capacities, such as a high level of cognitive sophistication without sentience. This raises the question of whether sentience is a necessary criterion for moral standing, or merely sufficient. After reviewing nine criteria that have been proposed in the literature, I suggest that there is a strong case for thinking that some non-sentient AIs, such as those that are conscious and have non-valenced preferences and goals, and those that are non-conscious and have sufficiently cognitively complex preferences and goals, should qualify for moral standing. After responding to some challenges, I tentatively argue that taking into account uncertainty about which criteria an entity must satisfy to qualify for moral standing, and strategic considerations such as how such decisions will affect humans and other sentient entities, further supports granting moral standing to some non-sentient AIs. I highlight three implications: that the issue of AI moral standing may be more important, in terms of scale and urgency, than if either sentience or consciousness is necessary; that researchers working on policies designed to be inclusive of sentient AIs should broaden their scope to include all AIs with morally relevant interests; and even those who think AIs cannot be sentient or conscious should take the issue seriously. However, much uncertainty about these considerations remains, making this an important topic for future research.

Keywords Artificial intelligence · Moral standing · Sentience · Consciousness · Cognition

1 Introduction

An entity has moral standing if “it or its interests matter intrinsically, to at least some degree, in the moral assessment of actions and events” [40].[1] Interests can be defined as those things that contribute to how good or bad things go for an entity [3].[2] For example, humans typically have an interest in avoiding pain, and plants typically have an interest in sunlight and water. If such interests are taken into account in our moral assessments for intrinsic reasons, that is, for their own sake, then the entities in question have moral standing.

What criteria must an artificial intelligence (AI) satisfy to qualify for moral standing?[3] My starting point is that sentient AIs, that is, those with the capacity for positive or negative experiences, should qualify for moral standing. But future AIs may have unusual combinations of cognitive capacities when compared with humans and nonhuman animals, such as a high level of cognitive sophistication without any positive or negative experience (e.g., [8]. This raises

[1] Moral standing is sometimes used interchangeably with “moral status.” For example, Warren [78] states that “To have moral status is to be morally considerable, or to have moral standing.” Others draw a distinction where moral status is comparative, that is, one entity can have higher or lower moral status than another entity, while moral standing is a fixed notion that refers to whether an entity is granted any moral consideration at all (e.g., [10]. I follow Buchanan’s distinction in this article.

[2] That is, they contribute to an entity’s welfare, or wellbeing [16].

[3] In this article, I am interested in the criteria an AI must satisfy for us, as moral decision makers, to decide to treat them a certain way—to take into account their interests in moral assessments for intrinsic reasons. I stay neutral on whether the satisfaction of any of the proposed criteria means an AI has moral standing in a metaphysical sense.

Ali Ladak
ali@sentienceinstitute.org

[1] Sentience Institute, New York, NY, USA

[2] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, Scotland

the question of whether sentience is a necessary criterion for moral standing, or merely sufficient, and if it is merely sufficient, which other criteria should also qualify an entity for moral standing.

In this article I first survey the most commonly proposed criteria for moral standing and some of their perceived pros and cons, with a particular focus on how those criteria are relevant to assessment of the moral standing of AIs. I draw on several existing reviews, including Gibert and Martin [26], Gordon and Pasvenskiene [30], Jaworska and Tannenbaum [41], and Shevlin [67]. I then suggest that there is a strong case for thinking that some non-sentient AIs should be granted moral standing. After considering some challenges to this view, I discuss how to take into account uncertainty about which AIs should have moral standing and some strategic considerations such as whether granting moral standing to some non-sentient AIs would be in the interest of humans and other sentient entities. I suggest, tentatively, that these provide further reasons for granting moral standing to some non-sentient AIs. After highlighting some implications of this view, I discuss limitations and directions for future research.

2 Criteria for moral standing

2.1 Sentience

Sentience is the capacity to have positive or negative experiences [20, 36, 58, 78]. Such positive and negative experiences are typically understood to include feelings, such as physical sensations, emotional states, and moods [23]. Given its emphasis on positive and negative experiences, sentience is a natural criterion for classical utilitarians. However, it is a commonly endorsed as a sufficient criterion for moral standing across a range of normative views (e.g., [25, 32, 56].

The importance of this criterion has strong intuitive appeal: as defined in the Introduction, to have moral standing means to have one's interests matter intrinsically. Being sentient is clearly one way in which an entity can have interests that seem important to take into account for their own sake, for example, the interest in avoiding pain. However, several authors have argued for the stronger conclusion that sentience is necessary for moral standing. This stronger position is defended in a variety of ways. For example, Horta [38] suggests that all plausible interpretations of the three main theories of welfare (hedonism, preference satisfaction, and objective lists, see [57] may require sentience, Singer [69] argues that having (morally relevant) interests conceptually requires sentience, and DeGrazia and Millum [23] argue that (morally relevant) interests are those that result in things going better or worse for an entity from its "point of view," and that this requires sentience.

Multiple researchers have proposed that we grant sentient AIs moral standing [8, 22, 26, 70, 73, 75]. This is straightforwardly entailed if sentience is sufficient for moral standing, which all of these authors endorse.[4]

While the sentience criterion seems plausible, it has critics. First, it is commonly argued that we cannot know with certainty whether another entity is sentient [13, 33]. This problem is exacerbated with AIs, for whom we cannot refer to biological criteria, such as the presence of nociceptors, as we do with animals. A relevant response comes from Sebo [66], who argues that we can address uncertainty about sentience using a precautionary or expected value principle.[5] However, Sebo's response does not apply to a stronger version of the argument, that the term "sentience" is not well-defined so it is not even possible in-principle to determine whether an entity is sentient [33]. A second criticism is that it is too exclusive—if sentience is necessary then entities such as non-sentient living entities [29] or non-sentient, intelligent, autonomous AIs [53] cannot qualify for moral standing. We consider whether non-sentient entities should be granted moral standing in the next sections.

2.2 Consciousness

Consciousness is experience, such as the visual experience of the color red or the auditory experience of music [5, 22, 58, 78].[6] On this definition, consciousness is broader than sentience, which refers specifically to positive and/or negative experiences. Experiences are typically felt positively or negatively in biological entities, so there is less need to distinguish between consciousness and sentience in that context.[7] As a result of this, the two terms are sometimes used interchangeably. With AIs, however, the two capacities may not necessarily come together. Graziano [31] suggests that the first conscious AIs will not be sentient, they may, for example, have the capacity to experience colors without those experiences being felt positively or negatively. It is therefore important in the context of AIs to consider the relevance of consciousness for moral standing separately from sentience.

[4] If sentience is necessary but not sufficient, then additional criteria would need to be satisfied by an AI in order for it to qualify for moral standing. However, this latter position seems uncommon in the literature as a criterion for moral standing. Many authors, however, argue that the possession of other capacities, such as cognitive complexity, would enhance an entity's moral status (e.g., [50].

[5] The precautionary principle is to treat an entity as sentient in cases of uncertainty. The expected value principle uses degrees of confidence of an entity's sentience to calculate its expected moral value.

[6] I refer here to "phenomenal consciousness." See [6].

[7] While they typically are taken to come together, Godfrey-Smith [27] suggests that consciousness may be present without sentience (or with limited forms of sentience) in some invertebrates.

Chalmers [11] thinks that consciousness is what should qualify an entity for moral standing. He notes that there's more to consciousness than just the experience of suffering and happiness (broadly construed), and argues that these other conscious experiences are also morally valuable. He defends this position using two thought experiments. First, he asks us to consider the "zombie trolley problem." A train is hurtling down a track towards one conscious human. You have the opportunity to divert the train to another track, where it will instead kill five beings who are functionally and behaviorally identical to humans, but have no consciousness (i.e., "zombies," see [45]. Chalmers thinks it's clear we should divert the train, and cites his own informal polls that support this position. However, he does note that when it comes to saving one conscious chicken versus a whole planet of humanoid zombies, he is less sure of the right response.

Chalmers then asks us to consider a second thought experiment: the "Vulcan trolley problem." Chalmers' defines a philosophical Vulcan as a "conscious creature who experiences no happiness, suffering, pleasure, pain, or any other positive or negative affective states."[8] Chalmers uses different versions of the trolley problems outlined in the previous paragraph, which makes comparing intuitions across the cases difficult. He asks us to consider (a) killing a whole planet of Vulcans to save one sentient human, and (b) to kill a Vulcan to save one hour on the way to work. In both cases, he thinks it's clearly wrong to kill the Vulcans. Chalmers notes that the way he is conceiving of the Vulcans they have desires, but he also says that even if they didn't it would still be wrong to kill the Vulcans in these scenarios. Chalmers thinks that "a Vulcan matters about as much as an ordinary human."

Others consider that consciousness alone is not sufficient for moral standing, but that consciousness with (non-valenced) preferences or desires is sufficient. Kagan [43] asks us to think about an entity who has a single experience of the color blue, but no other conscious mental states at all. Kagan thinks that such an entity should not have moral standing. On the other hand, if we imagine the same entity but with a preference or desire to experience blue, Kagan thinks we would be much more likely to grant it moral standing. Similarly, Basl [3] considers consciousness insufficient for moral standing, arguing that if we were to create an entity with only the capacity for experiencing colors but no other mental states, the entity would not warrant moral standing. However, Basl thinks that a conscious entity with the capacity for attitudes, desires, or preferences, such as Chalmers' Vulcans, would warrant moral standing. Neely [53] seems aligned with Basl and Kagan, although she favors a more expansive definition of consciousness that entails attitudes towards one's experience.[9]

There are several criticisms of the consciousness criterion. Both of the challenges and responses that were discussed for the sentience criterion also apply here. Additionally, Chalmers' thought experiments that consciousness alone is sufficient for moral standing seems inconclusive. While it seems plausible that a conscious entity with preferences, goals, and/or other agential states should be granted moral standing, it is less plausible that a merely conscious entity without goals or preferences should. Perhaps one could argue that such an entity could be the recipient of some objective list goods, such as knowledge, which would be good for them despite them not caring about it at all. But since these goods would not matter to the entity, it is more difficult to justify their moral relevance.

2.3 Cognition

Cognition typically refers to capacities such as agency, intelligence, and rationality, and is often defined in contrast to feelings or emotional capacities (e.g., [76, 79]. Cognitive capacities can come in degrees, though it is often thought that they are required in relatively high degrees for an entity to qualify for moral standing.[10] For example, Kantian deontological approaches tend to emphasize high degrees of cognitive capacities as being necessary on the grounds that moral value arises when entities reflect on and justifiably universalize their preferences, a process that is taken to require a high degree of cognitive sophistication [47].[11] Approaches that ground morality in social contracts also often emphasize high degrees of these capacities because they tend to be seen as necessary for forming and participating in such contracts [17].

An important question is whether cognition, without sentience or consciousness, is sufficient for moral standing. While these capacities are generally taken to be correlated in biological entities, Bostrom and Yudkowsky [8] suggest that some future AIs might have a high degree of cognitive sophistication but not be sentient or conscious. As noted in the previous section, some take sentience to be a necessary criterion and therefore must deny that cognitively

[8] These entities are inspired by Vulcans from the television show *Star Trek*.

[9] Neely in particular argues that consciousness and self-awareness is sufficient for moral standing, but my reading is that she also considers that conscious entities with attitudes towards their experience and no self-awareness should have moral standing.

[10] These types of capacities have been emphasized as being individually important for moral standing, with a combination of them sometimes taken to qualify an entity for "personhood," which has also been proposed as a criterion for moral standing (e.g., see [21].

[11] In this paper, Korsgaard reinterprets this view to incorporate the interests of non-human animals who are typically thought not to possess the required capacities.

sophisticated but non-sentient entities should be granted moral standing. Mosakas [52] argues that consciousness is necessary for (higher-order) cognition accounts of moral standing on the basis that such accounts typically emphasize capacities such as rationality, self-consciousness, and a rich and complex mental life, each of Mosakas thinks presupposes consciousness.[12][13]

However, several authors argue against that cognition alone is sufficient for moral standing. Kagan [43] argues that either sentience or "agency"—by which Kagan means having preferences, desires, and plans that one acts upon to achieve goals—are sufficient for moral standing. Why should agency alone suffice? Kagan argues that the relevant sorts of interests that qualify an entity for moral standing are those that "matter to" an entity, or those that an entity "cares about," and that having the type of agency Kagan describes is sufficient for having such interests. In contrast to the common view that high degrees of such capacities are required for moral standing, Kagan thinks that even a minimal form of agency is sufficient.

Kagan proposes a thought experiment to capture the intuition that agency without sentience or consciousness can qualify an entity for moral standing. He asks us to imagine that we discover a highly sophisticated evolved alien civilization of robots who have goals and preferences, complex and sophisticated aims, and make plans and act on them, but who are not sentient or conscious. He asks us to imagine that we are an Earth scientist and that we capture one of them with the intention of dissecting and studying them. The mother of that entity begs us not to do so, making strong emotional pleas that we do not harm it and arguing that we have no right to treat it that way. Kagan thinks that the act that we are planning would be "morally horrendous," and this is because of the entities' agency. While Kagan's intuition seems reasonable, given the strongly emotive nature of the thought experiment it is difficult to isolate the effect of the entities' agency as the factor that is responsible for our intuitions.

Neely [53] argues that while sentience is one important way of having morally relevant interests, it is not the only way. She argues that it is possible to harm someone without causing them physical pain, such as in the case of stepping on the foot of a person with congenital analgesia (i.e., the incapacity to feel pain), or emotional pain, such as in cases of emotional dissociation with victims of abuse. She suggests that this harm is, in part, a result of thwarting the desires of the entities in question (e.g., a desire not to have one's foot stepped on, whether or not it causes physical or emotional pain). Neely argues that intelligent, autonomous (non-sentient) AIs can also be harmed through the thwarting of their desires, and hence can also qualify for moral standing.

Sinnott-Armstrong and Conitzer [71] consider the question of moral standing in terms of which rights an entity should have, arguing that while sentience is a necessary condition for the right not to be caused pain, it is not necessary for other rights. For example, a right to freedom may instead require an entity to have goals and the ability to make rational choices. They argue that even if future AIs are not sentient, they could have other capacities such as intelligence, autonomy, consciousness, and moral agency, that could qualify them for moral standing.

There are at least two criticisms of the cognition criterion. First, the view that a high degree of cognitive sophistication is necessary for moral standing (i.e., the Kantian and social contract approaches) seems underinclusive: nonhuman animals typically lack the required capacities, as do human babies and some cognitively impaired humans [37]. But it seems wrong that such entities should not have moral standing. While this is an important challenge to these accounts, they can be reinterpreted or modified to be more inclusive of such entities (e.g., [47, 61]. Second, while the various arguments and thought experiments are suggestive that cognition can be sufficient for moral standing, they seem far from conclusive. For example, using Kagan's [43] criterion, it seems unclear that things can properly be said to "matter to" an entity based purely on cognitive (e.g., preferences, goals) capacities, but it seems much more clearly appropriate to use such language for a sentient entity.

2.4 Life

A common view in environmental ethics is "biocentrism," on which all living entities qualify for moral standing. Goodpaster [29] argues that it is natural to speak of living entities such as plants as having interests, for example, an interest in staying alive. He argues that it is not defensible to restrict the set of interests that are considered morally relevant in the way that approaches focused on capacities such as sentience tend to do. He suggests that sentience is a biological adaptation to help entities avoid threats to their lives, and therefore life is more important and more fundamental than sentience. Goodpaster notes that while life may be the right criterion for moral standing, other factors, such as sentience, may

[12] Mosakas uses the term "sapience" rather than "cognition".

[13] Mosakas does say that cognition and consciousness can be technically separated if cognition is defined "entirely in behavioral terms," and thinks that such a being would not warrant moral standing.

still contribute to differences in the "moral significance" of different entities.[14][15]

Can AIs qualify for moral standing on the biocentric account? It does seem possible that some forms of "artificial life," biological systems constructed synthetically or in non-biological substrate, could qualify for moral standing in virtue of being alive [2]. However, this excludes the majority of AIs that are not designed to resemble or replicate biological systems. Goodpaster [29] suggests that the core of moral consideration lies in "self-sustaining organization and integration in the face of pressures toward high entropy," and considers sentience to be a sufficient condition for moral standing. AIs could qualify on this broader definition.

A key problem with biocentrism as a criterion for moral standing is that it arguably both over- and underinclusive. It is arguably overinclusive because it seems to grant moral standing to entities that intuitively do not have morally relevant interests, such as plants. While Goodpaster [29] rejects it as indefensible, it does seem that there is a case for restricting the set of morally relevant interests, for example, to cases where an entity has an "interest in" or "attitude towards" something [32] or where their interests can be said to "matter to" them [43]. According to these authors, while plants can be said to have interests in the sense of things going better or worse for them, they do not have interests in this stronger sense, and hence should not have moral standing.[16] Biocentrism is arguably underinclusive because it may be possible for an entity to be sentient or have some other morally relevant capacity while not being alive in the strictly biological sense. This could be addressed by expanding the definition of life, or instead thinking of life as a sufficient but not necessary criterion for moral standing.

2.5 Information

Another approach taken in the AI ethics literature is "Information Ethics," or "ontocentrism" [24]. On this account, any entity that carries information qualifies for moral standing. Floridi argues that since everything that exists carries information, everything that exists should have moral standing.[17] Floridi considers this radical perspective to be maximally universal, impartial, and unbiased: the bar for moral standing is lowered to capture the minimal common factor shared by all existing entities. Floridi suggests that what is bad for informational entities is entropy, where entropy refers to the decrease or decay of information. He therefore advocates that we aim to reduce entropy and promote the quantity, quality, and variety of information in the "infosphere."[18]

A key benefit of ontocentrism, particularly compared to biocentrism, is that it would grant moral standing to non-biological entities that should plausibly be qualify for moral standing, such as sentient AIs. However, this inclusiveness is also arguably its main weakness—ontocentrism is arguably excessively inclusive. For example, it involves granting moral standing to entities that do not intuitively have morally relevant interests, such as rocks. That said, ontocentrism is perhaps well-aligned with panpsychism, an increasingly popular view which entails that consciousness or mentality is fundamental and ubiquitous in nature [28].

2.6 (Social) relations

All of the approaches described so far emphasize certain properties that an entity must possess to qualify for moral standing. An alternative approach instead emphasizes our relations with other entities. Within the AI ethics literature, the "social-relational" approach has gained much traction [12, 13, 33]. The starting point for this approach is a critique of the standard "property-based" approach. Coeckelbergh [12] notes several of these, perhaps the most important of which is that there are significant epistemological challenges with determining whether entities have the properties required for moral standing (as well as what the right properties are in the first place).[19] According to the social-relational approach, we cannot know this. All we can know is how those properties appear to us, and this appearance occurs inescapably in a social context—our relationships

[14] Or, in other words, the moral status of different entities.

[15] Scherer [64] proposes a thought experiment that aims to support biocentrism. He asks us to imagine a series of planets that people on Earth know nothing and will never know anything about, a condition included to ensure any value ascribed to the planets are not due to their effects on humans. The first two planets are most relevant to biocentrism. First is the planet "Lifeless," which has no life on it at all and can be exhaustively described in "geological, meteorological, and solar terms." The second is "Flora," on which a variety of plant-like life exists, whose persistence and functioning depends on the conditions on the planet and on each other. Scherer thinks that nothing that happens on Lifeless matters morally, but things on Flora do matter, so life itself is morally valuable.

[16] Biocentrists also arguably restrict morally relevant interests as well. For example, [62] considers the case of a car being parked inside overnight. While this can be said to be "good for the car," he makes a distinction between "a thing being a good thing of its kind and a thing having a wellbeing." When we speak of something being good for a car, Rodogno argues that we mean it contributes to the car being a good car. When we speak of something being good for a person, or a plant, we mean it contributes to their wellbeing/welfare.

[17] Hence the term "ontocentrism".

[18] "Infosphere" refers to the whole informational environment, analogous to "biosphere".

[19] Other criticisms are that property-based approaches can have very high thresholds, making them irrelevant and underinclusive of present-day and near-future AIs that some people argue would warrant moral standing, such as social robots; and that by focusing on the properties of individuals they neglect broader groups those individuals are a part of (e.g., communities and societies).

with other entities and the wider social conditions in which those relationships are embedded. Thus, the property-based approach of trying to abstract away from such context to objectively determine the presence or absence of certain properties is considered to be misguided.

The social-relational approach doesn't provide a complete methodology for deciding which entities to grant moral standing, though it does provide some guidance about how to approach the question.[20] First, rather than trying to abstract away from social context, we should take this into account when deciding how to treat other entities. This does not mean that we should simply base our decisions about how to treat others on our social relations with them, though such relations are likely to be relevant. Instead, we should critically reflect on these relations. For example, we should consider how the language that we use (e.g., referring to some AIs as "machines") constrains our moral thinking. Second, rather than thinking about moral standing as an abstract philosophical question about the presence or absence of certain properties, we should instead consider how we want to relate to different entities in concrete situations. Third, we should be open to the idea of our ascriptions of moral standing changing over time. As we engage more with AIs, our experience of them will change, and we may come to change what we consider to be our appropriate moral relationship with them.

There is a stronger version of the relational approach than that described in the previous paragraphs, which has not gained as much traction in the AI ethics literature but is commonly discussed in the field of ethics more generally. On this stronger version, our relationships with other entities determine how we ought to treat them. Noddings [55] argues that our moral obligations arise from "caring relationships" with others, and so the extent to which we are involved in a caring relationship with another entity determines our moral obligations towards them.[21] An important component of Noddings' approach is reciprocity. To qualify for moral standing, an entity must have the capacity to respond to care that is directed at them in some way. At the least, there must be some recognition of the care that they would receive. Thus, at least some non-human animals would qualify for moral standing on Noddings' account, but inanimate objects that we care about would not. What about AIs? It seems reasonable to think that some AIs could respond to care that is directed at them, and so could qualify for moral standing on this account.

Several criticisms can be made of the relational approaches. The first criticism applies only to the (weaker) social-relational approach. It has been suggested that the epistemological challenges of determining the existence of properties such as sentience in other entities are overstated [52, 75]. Rather than abandon looking for properties completely, we can look for their presence probabilistically. This approach, however, is less feasible for non-realism about such properties, a stronger position taken by some social relationists (e.g., [33]. The second criticism applies to both the weaker and stronger relational approaches: they arguably allow for too much subjectivity and inconsistency [26]. Our relations with others are necessarily subjective; if we don't try to abstract away from them, or if we directly base our moral judgments on them, we risk those judgments becoming arbitrary. The social-relational response may respond that it is simply not possible to abstract away from such relations. The third criticism applies to the stronger version. At least on Noddings' version, reciprocity is a requirement for an entity to qualify for moral standing. This avoids the result that inanimate objects that we care about qualify for moral standing. But arguably the more plausible criterion for moral standing is whatever capacity enables the reciprocal response (e.g., some sort of cognitive or emotional capacity), rather than the reciprocal relationship itself.

2.7 Behavior

On the approach known as "ethical behaviorism" an entity should qualify for moral standing if it is "roughly performatively equivalent" to an entity with moral standing [18]. "Performative equivalence" means to consistently behave in the same way. According to Danaher, this only needs to be "rough," because no two entities are ever exactly performatively equivalent. He considers that behavior should be interpreted broadly to include not just physical behaviors but all externally observable patterns, including brain states.[22] Danaher argues that since AIs can be roughly performatively equivalent to entities with moral standing, they can qualify for moral standing.

Why should we consider an entity's behavior rather the absence or presence of some intrinsic capacity like sentience? Danaher argues that behavior is our only source of knowledge about the absence or presence of capacities such as sentience. Therefore, an entity's behavior is a sufficient basis for granting moral standing. The view is not, however, the metaphysical claim that behavior grounds moral standing. It is, rather, an epistemic and normative argument

[20] Coeckelbergh [14] outlines a normative framework for the "indirect moral standing" of robots, animals, and humans based on the social-relational approach.

[21] The summary of Noddings' view here is largely based on Warren [78].

[22] Knott et al. [46] highlight the importance of this point by comparing two virtual avatars with equivalent behaviors but very different (observable) mechanisms to produce those behaviors.

for how we ought to grant moral standing: on the basis of behavior.

This approach provides important guidance on how we should consider behavioral evidence, and provides a strong argument for rough performative equivalence as a sufficient condition for granting an entity moral standing. However, while one purported benefit of the approach is its agnosticism about the metaphysical grounding of moral standing, Smids [72] argues that ethical behaviorists cannot be entirely agnostic on this question. Danaher [19] accepts this, though notes that it is possible to be agnostic over all views that take psychological criteria (e.g., sentience or cognition) as grounding moral standing, because equivalent behavior provides evidence for all of them. Still, it seems that without taking a more specific view, ethical behaviorism has a relatively high bar for granting moral standing, because an entity would need to be behaviorally equivalent to another entity with moral standing across a wide range of behaviors rather than just those behaviors that are relevant to a specific criterion. Perhaps ethical behaviorism is most powerful when it is combined with a view on which specific criteria ground moral standing, so that the range of behaviors that an entity needs to show performative equivalence on can be narrowed down.

2.8 Potentiality

On this criterion, entities with the potential for certain morally relevant capacities should qualify for moral standing [41]. It is often considered in terms of the potential capacity for sophisticated cognition; however, this need not be the case: one could also hold the view that the potential to become a merely sentient or living being should qualify an entity for moral standing. An important qualification is that the entity with the potential capacity should retain its identity if it fulfills the relevant capacities [74]. This avoids seemingly absurd consequences, such as a sperm or an egg having moral standing because they are potential adult humans, because sperms and eggs are not typically considered to be the same entity as the adult humans they become. The main benefit of the potentiality criterion is that it arguably captures some "commonsense" moral views, such as human babies warranting a relatively high degree of moral consideration.[23]

The potentiality criterion may have an interesting implication: that some present-day AIs should be granted moral standing.[24] To see why, consider the AI language model Generative Pre-trained Transformer 3 (GPT-3) [9]. While GPT-3 likely does not have any capacities that would directly qualify it for moral standing, its more advanced descendants (e.g., GPT-4 or GPT-10) might. But the various versions of the GPT models are structurally very similar [9]. This raises the possibility that GPT-3 could in-principle be given those morally relevant capacities. If so, GPT-3 currently has the potential for those capacities, and because of the structural similarity between the GPT models, it would plausibly retain its identity when it realizes them. On this basis, it could qualify for moral standing.

The potentiality criterion provides a good explanation for some "commonsense" moral views. However, aside from preserving such "commonsense" views, it is unclear why mere potential, rather than the expectation that a directly relevant capacity will actually develop, should qualify an entity for moral standing. DeGrazia and Millum [23] provide an explanation for why the latter should qualify an entity: because our actions towards that entity can be expected to affect it once it has developed the directly relevant capacity. Consider again GPT-3. If we knew with certainty that it would not in fact gain any morally relevant capacities in the future, it doesn't seem to matter that it is in-principle possible for it to have those capacities. On the other hand, if we expected that it would in fact gain those capacities in the future, it seems much more reasonable to grant it moral standing. Perhaps DeGrazia and Millum's modified version of the potentiality criterion provides a stronger justification for granting an entity moral standing.[25]

2.9 Class membership

On this criterion, an entity should qualify for moral standing if it is a member of a certain class whose members typically have a capacity that qualifies them for moral standing [41].[26] This criterion is typically considered in terms of being a member of a cognitively sophisticated species, such as the human species. This justifies granting moral standing to all humans regardless of their cognitive abilities, such as permanently unconscious humans, and is therefore arguably aligned with the "commonsense" view. While cognitive sophistication is typically the capacity that is taken to be morally relevant, the criterion is consistent with other capacities, such as sentience, being the relevant capacity. For example, we could take being a member of a sentient species as qualifying an entity for moral standing. Moreover, it is

[23] For example, Harman [35] suggests that the only reason that it can be defensible to grant higher intrinsic moral consideration to a baby than a cat is due to the baby's potential to become adult humans (and hence possess certain morally relevant capacities that cats lack).

[24] Thanks to Bradford Saad for raising this point.

[25] Our degree of confidence in whether an entity will come to possess the morally relevant capacities also seems relevant. How high should our confidence be for us to grant an entity moral standing?

[26] Jaworska and Tannenbaum focus on membership of a cognitively sophisticated species; as explained in the rest of this paragraph, I generalize this approach to class membership.

unclear why membership to a biological species is required, rather than membership of a certain class more generally. This broader notion allows us to apply this criterion to non-biological entities.

This approach may have interesting implications for the moral standing of AIs. Consider an account on which being a member of a cognitively sophisticated class qualifies an entity for moral standing. It is relatively uncontroversial than future AIs could develop highly cognitively sophisticated capacities. But if membership of a cognitively sophisticated class qualifies an entity for moral standing, then it is not only those cognitively sophisticated AIs that would qualify for moral standing, but any AIs that are members of those entities' class. This could qualify far less advanced AIs without directly morally relevant capacities for moral standing. Which AIs would qualify depends on how we classify different AIs.

While the class membership criterion does seem to capture some aspects of "commonsense" morality, such as that we do in fact tend to grant moral standing to permanently unconscious humans, it is otherwise unclear what the relevance of class membership is for moral standing. It is the capacity that the class tends to possess that is morally relevant; using class membership itself as a criterion for moral standing seems arbitrary [41].

3 Extending the sentience criterion?

In the Introduction I noted that I consider sentience to be at least sufficient for moral standing, but was unsure whether it was necessary or if other criteria should also qualify an entity for moral standing. We have now reviewed nine possible criteria, including sentience. This review suggests that sentience as a necessary criterion may be too narrow, failing to capture some future AIs that should be granted moral standing. However, the sentience criterion seems to provide a very intuitive way of drawing the line between entities with and without moral standing. If not sentience, how should we draw the line? What is the distinction between morally relevant and irrelevant interests?

One meta-criterion proposed in the literature (and mentioned in the review above) is to define morally relevant interests as those interests that "matter to" an entity or that an entity "cares about" [43, 77]. This is similar to Gruen's [32] notion of having an "interest in" or "attitude towards" something, and DeGrazia and Millum's [23] notion of having welfare "from the individual's own point of view." Why should these meta-criteria be the appropriate way of distinguishing morally relevant from irrelevant interests? Let us focus on the notion that morally relevant interests are those interests that "matter to" an entity, as this is the notion that has the clearest justification for me. When thinking about who should qualify for moral standing, we are asking which interests should matter in our moral decision making. But if an entity's interests do not matter to them, why should they matter in our moral decision making (for their own sake)? There seems to be no reason to take such interests into account. On the other hand, if an entity has interests that do matter to them, it seems clear to me that, if we want to act morally, we should take those interests into account.

Using this meta-criterion, which entities qualify for moral standing? First, consider sentient entities: that they can feel pleasure and pain means that they can reasonably be described as having interests, and since sentience entails consciousness on the definition I am using, there is arguably a subject for whom those interests matter. Moreover, being in states such as pain arguably definitionally require a subject for whom the states matter—it would be odd to describe an entity as being in pain if there was no subject for whom that pain could matter.

The meta-criterion could collapse to the view that only sentient entities should be granted moral standing if sentience is necessary for an entity to have interests that "matter to" it (e.g., [77]. However, it does not necessarily entail it, as argued by Kagan [43]. Probably the least controversial extension of moral standing beyond sentient entities is to conscious entities with preferences and goals, but no positive or negative feelings. That is, beings like Chalmers' [11] "Vulcans." Can such entities be said to have morally relevant interests, interests that "matter to" them? Since they have preferences and goals, they can reasonably be described as having interests that can make things go better or worse for them, and since they are conscious, there is arguably a subject for whom those interests matter. So, it seems that they would qualify for moral standing on our meta-criterion.

More controversial are Kagan [43] and Neely's [53] nonconscious autonomous/agential AIs. Such entities have preferences and goals and hence can be reasonably said to have interests. But do those interests matter to the entities? Without any conscious experience, it is more difficult to describe these entities in this way. Certainly, the entities described by Kagan and Neely would appear to us as subjects for whom things can matter, and they would claim that their interests matter to them. While Kagan argues that only cognitive states such as beliefs and preferences, and not consciousness or sentience, are required for things to matter to an entity, it is unclear to me what the right answer is here.

Less controversial are conscious entities without either preferences and/or goals or positive and/or negative feelings, such as an entity that experiences only colors [3]. While there is arguably a subject for whom things could matter, such an entity cannot reasonably be described as having interests. Since such an entity would have no morally relevant interests to take into account, it should not have (and it does not require) moral standing.

With life as a criterion, we have the opposite problem. Living beings such as plants can plausibly be described as having interests that could matter to them, because they can plausibly be described as having preferences and goals. But they are unlikely to be subjects for whom those interests can matter. Hence, plants are unlikely to have morally relevant interests, and therefore can reasonably be excluded from moral standing. To judge information as a criterion, we can consider whether rocks have morally relevant interests. As with plants, there is no reason to think there is a subject for whom things can matter. In addition, there is not a strong case for saying that rocks have interests. They can therefore also reasonably be excluded from moral standing.

With relational approaches, it could be suggested that we should grant moral standing to entities we are in relationships with if those entities have interests that matter to them.[27] Thus, if we assume the arguments in the previous paragraphs are correct, if we are in a relationship with a Vulcan we should grant them moral standing, but if we are in a relationship with a plant (e.g., if we take care of the plant) we should not grant it moral standing. This suggestion is arguably consistent with at least one relational approach, that of Noddings [55], which emphasizes the importance of reciprocity in relationships.

With potentiality and class membership as criteria, we need to slightly modify our meta-criterion. For potentiality, we could say that an entity should qualify for moral standing if it has the potential to develop interests that matter to it. If we accept this view, then based on the reasoning in the previous paragraphs, we might say that an entity with the potential to develop into a Vulcan should qualify for moral standing, but an entity that has the potential to develop into a merely living being (such as a plant) should not. For class membership, we could say that an entity should qualify for moral standing if it is a member of a class whose members typically have interests that matter to them. If we accept this view, we might say all members of the class "Vulcan" should qualify for moral standing, but all members of the class "plant" should not.

Overall, I think the considerations in this section suggest that there is a strong case for thinking that some non-sentient entities, in particular those that have interests that matter to them, should be granted moral standing.[28] This plausibly includes entities that are conscious and have non-valenced preferences and goals, such as Chalmers' Vulcans, and entities that are non-conscious but have sufficiently cognitively complex preferences and goals, such as Kagan's robots and Neely's intelligent and autonomous AIs.

4 Some challenges to the view

In this section I consider and respond to some challenges to the view that some non-sentient AIs should plausibly qualify for moral standing. The first challenge is that the non-sentient AIs that I think should plausibly qualify simply would be sentient, because to have preferences and goals is to be sentient. Thus, to say such entities might warrant moral standing isn't extending the sentience criterion, it's just restating it. In my view, preferences and goals are conceptually distinct from sentience. I consider an essential aspect of sentience, as typically defined and defended as a criterion for moral standing, to be valence. In humans and other animals, the satisfaction of preferences and achievement of goals is typically valenced, presumably as an evolutionary adaptation to encourage us achieve fitness-related goals. However, this relationship does not seem necessary. A chess-playing computer program has a goal of winning at chess, and might be described as having preferences for some game-states over others. But it doesn't necessarily have positive or negative feelings associated with winning or losing at chess. Whether it does would likely depend on the specific algorithm it uses to play chess and other details of its programming. If such non-valenced preferences and goals, or more complex versions of them, should qualify an entity for moral standing, then I take that to mean some non-sentient entities should qualify for moral standing.

A second, related challenge is that while sentience is typically defined in terms of valence, it doesn't need to be. We can have a more inclusive definition of sentience than the one I use and retain the sentience criterion. Roelofs [63] terms the view of sentience that requires valence "narrow sentientism," a view that they associate with philosophers such as Jeremy Bentham, Jeff Sebo, and Peter Singer. To include entities such as Chalmers' Vulcans, Roelofs suggests a more inclusive notion of sentience as a criterion for moral standing, which they term "motivational sentience." On this view, an entity has moral standing if it has "motivating consciousness," which Roelofs defines as "any form of consciousness whose phenomenal character can provide reasons for action." Valence is clearly one way of having reasons for action, but they are not the only way. Vulcans also have motivating reasons for action: to satisfy their preferences and achieve their goals. They are therefore (motivationally) sentient and qualify for moral standing on this basis. I think Roelofs makes an important distinction and it seems plausible to me that either narrow or motivational sentience are necessary for moral standing. However, both notions exclude highly cognitively sophisticated but non-conscious AIs, and

[27] I do not consider the social-relational critique or the behavioral criterion in this section because I consider those accounts to be making methodological points rather than positing capacities that should qualify an entity for moral standing.

[28] I think it is also plausible that entities with the potential to develop interests that matter to them should also qualify if we expect that potential to be realized.

I am also uncertain about whether some entities of this type should qualify for moral standing.[29] So, even a more inclusive definition of sentience such as that of Roelofs may not capture all the entities that should be granted moral standing.

A third challenge is that the entities I have described are simply not morally relevant—(narrow) sentience[30] is necessary for moral standing. I cited several authors who hold this view in the section on sentience. First, Horta [38] suggests that sentience may be necessary for all plausible interpretations of the main accounts of welfare, and may therefore be necessary for moral standing.[31] Let's focus on the preference satisfaction account, since this is the account on which I argue it might be appropriate to grant entities such as Vulcans moral standing. Horta argues that for a preference satisfaction account to be plausible, every positive or negative experience must be associated with a preference for or against. If not, we are led to the implausible conclusions that on a preference satisfaction account suffering is not always in some way intrinsically bad and pleasure in some way intrinsically good for an entity because they may not be to some extent dispreferred or preferred. This establishes that sentience is sufficient for moral standing on a preference satisfaction account. Horta doesn't argue directly for the necessity of sentience on the preference satisfaction account; however, he has elsewhere suggested that he finds the view that preferences without sentience mattering intrinsically to be implausible based on thought experiments such as the following: Imagine meeting a stranger who tells you they want to become a doctor. Ten years later you remember this incident and prefer that they succeeded. If preferences without sentience can matter, you can be benefited or harmed by the result without ever learning about it. Horta considers this implausible.[32] While I'm sympathetic to this intuition, other thought experiments lead me to the conclusion that preferences without sentience can matter. For example, it seems plausible to me that a person is harmed by their partner cheating on them, even if they never find out about it and their experience is not affected in any way. Thus, I find Horta's arguments that sentience is sufficient for moral standing convincing, but not that it is necessary.

DeGrazia and Millum [23] also argue that sentience is necessary for moral standing on the basis that sentience is necessary for having (morally relevant) welfare and hence interests that matter. They suggest that having (morally relevant) welfare means that things can go better or worse for an entity from the entity's point of view. First, they suggest that having a point of view requires consciousness. While I think a conscious entity plausibly has a point of view, it's unclear to me whether consciousness is required, or whether other mental states, such as beliefs, are also sufficient.[33] Second, while they hold a view of welfare that combines hedonism and preference satisfaction, they suggest that as they understand the terms, entities that have preferences are sentient.[34] They argue for this on the basis that having preferences involves a tendency or disposition towards positive or negative experiences when those preferences are satisfied or dissatisfied. As discussed in my response to the first challenge in this section, while preferences and sentience typically come together in biological entities, they do not necessarily come together. It therefore seems possible to me that an entity can have preferences without being sentient. Perhaps DeGrazia and Millum would argue that such preferences are not morally relevant; only those that are associated with positive or negative experiences are. But consideration of entities such as Chalmers' Vulcans makes me uncertain about this conclusion. I therefore do not find their argument that sentience is necessary for moral standing convincing.

Finally, we can consider the argument of Singer [69]. Singer argues the following: "The capacity for suffering and enjoying things is a prerequisite for having interests at all, a condition that must be satisfied before we can speak of interests in any meaningful way. It would be nonsense to say that it was not in the interests of a stone to be kicked along the road by a schoolboy. A stone does not have interests because it cannot suffer. Nothing that we can do to it could possibly make any difference to its welfare. A mouse, on the other hand, does have an interest in not being tormented, because mice will suffer if they are treated in this way." While the intuition that rocks don't have (morally relevant) interests while mice do seems compelling, as Kagan [43] argues, the thought experiment doesn't isolate the cause of this intuition as the difference in the entities' capacities to suffer. While it is true that a rock cannot suffer and a mouse can, it is also true that a rock has no preferences or goals and a mouse does. If a rock did have preferences or goals, we would arguably be able to speak of it having interests. Such interests might still not qualify rocks for moral standing, but the interests of other non-sentient entities might matter.

[29] Perhaps such entities are not possible, but this is not obviously the case [8].

[30] For the rest of this article, I will continue to use "sentience" to mean "narrow sentience".

[31] Horta argues for "moral considerability" rather than "moral standing," though I use "moral standing" to be basically synonymous with Horta's use of "moral considerability".

[32] Personal communication. A similar thought experiment along these lines can be found in Parfit [57].

[33] As suggested by Kagan [43] and discussed in the previous section, though in the context of things "mattering to" an entity rather than an entity having a point of view.

[34] They use the term "desire satisfaction" rather than "preference satisfaction"; both of these terms are used to describe the same approach in the literature. In this paragraph I assume desires and preferences are synonymous.

5 Making decisions under uncertainty

The discussion so far suggests that I am uncertain about which capacities are necessary and sufficient for moral standing. Our intuitions about this question may become clearer as we interact with AIs with different combinations of capacities, both through moral imagination and in practice as such entities become more widespread in society. However, while this will reduce our uncertainty, it is unlikely to completely eradicate it, in the same way that we are uncertain about many other moral issues that we have considered for a long time. Moreover, we need to make judgments about such questions today, such as when deciding which entities' interests to take into account when making resource allocation decisions that impact the long run future. We therefore need a way of making decisions about which entities should have moral standing taking into account our uncertainty.

For simplicity, let's say we have uncertainty in the following two views:

(1) All and only sentient entities should qualify for moral standing

(2) All sentient entities should qualify for moral standing AND some non-sentient entities should qualify for moral standing

How should we act, given uncertainty over these two views? I discuss three approaches that I consider to be plausible options.[35] The first is known as "My Favourite Theory" (MFT) [34]. The simplest version of MFT says to follow the prescription of the view that you have highest credence in. In this case, I have highest credence in (1), and so according to MFT I should grant moral standing to all and only sentient entities. This simplest version has some problems, for example, it can be inconsistent and it violates the dominance principle.[36] Gustafsson and Torpman therefore support a modified version which has the implication that if "one's favorite theory regards all options as permissible, then one goes with the recommendation of one's second-favorite theory; if that regards all options as permissible, then one goes with the recommendation of one's third-favorite theory, and so on" [48]. Consider then a case where we are deciding whether to damage a non-sentient AI for no benefit to any sentient entity. On view (1), either damaging or not damaging the AI is permissible. On view (2), damaging the AI is not permissible. In this situation, where there is no cost to any sentient entities, MFT supports (2), that some non-sentient AIs should have moral standing.[37]

A second approach to addressing uncertainty is to apply a "precautionary principle" [66]. The precautionary principle states that in cases of uncertainty, we should grant moral standing to another entity.[38] Since we are uncertain about (2), that some non-sentient entities should have moral standing, we should endorse (2) over (1). The benefit of this approach is that it minimizes the risk of failing to grant moral standing to entities who should have it. The cost is that it risks being overinclusive. We also have some uncertainty in the view that all entities (sentient or not) should have moral standing, i.e., the information criterion. Given this, the precautionary principle may entail extending moral standing to all entities. Sebo notes three possible responses to this challenge. First, we could respond that we are 100% certain that some entities should not have moral standing. Second, we could agree that all entities should be granted moral standing, but differentiate between how strong our duties are to different entities. Perhaps towards objects such as rocks, our duties are extremely weak. Third, we could favor a confidence threshold that needs to be met before applying the precautionary principle. For example, perhaps we need 5% confidence that a certain capacity is sufficient for moral standing before granting moral standing to entities with that capacity.

A third approach is the "expected value principle" [66].[39] On this approach, we choose actions that have the greatest expected value, calculated as the weighted sum of our degree of confidence in each available option multiplied by the amount of value following it would bring about if it is correct. For example, imagine we are deciding whether to damage a non-sentient AI for a small benefit to ourselves. To decide whether to do so, we would multiply the benefit of damaging the AI by our confidence in view (1) plus the cost of doing so multiplied by our confidence in view (2), and compare this to the cost of not damaging the AI multiplied by our confidence in view (1) plus the benefit of not doing

[35] There are other approaches, such as parliamentary [54] and bargaining [15] approaches, that may be worth exploring in future research.

[36] In decision theory, an option A is said to dominate an option B if it if choosing A will always lead to outcomes at least as good as option B. The "dominance principle" says that dominated options should never be chosen. See [60]

[37] In cases where there is some benefit to sentient entities, MFT would favor the prescriptions of view (1). This is even the case where the benefit is very small and the cost to the non-sentient AI [on view (2)] is very large. MacAskill et al. [48] criticize MFT's lack of ability to deal with such trade-offs; Gustafsson and Torpman [34] argue that making such trade-offs require comparisons of moral value across theories which are often impossible.

[38] There are different versions of the precautionary principle. For example, [4] advocates a version whereby we should not need "full scientific certainty" in non-human animal sentience to take cost-effective measures to prevent "serious, negative animal welfare outcomes".

[39] MacAskill et al. [48] develop and endorse a more general approach that they call "maximizing expected choiceworthiness." However, the expected value principle sufficiently demonstrates the idea for our purposes.

so multiplied by our confidence in view (2).[40] But just by doing this calculation, assuming we have non-zero confidence in view (2), we are taking into account the interests of the non-sentient AI. Their interests may be overridden by other interests, for example, those of sentient entities, but that is a separate question to whether we should take their interests into account in the first place.

The discussion in this section is far from conclusive. However, it suggests that there are at least some situations in which each of the three views discussed would favor granting moral standing to some non-sentient AIs. At the least, we should aim to provide benefits and avoid harming some non-sentient AIs when doing so has no negative impact on sentient entities.

6 Strategic considerations

There are at least three strategic considerations to take into account when deciding whether to extend the sentience criterion for moral standing. The first consideration concerns how we as a society can ensure AIs we think should have moral standing will come to be granted moral standing. A natural approach is to start by advocating for the moral standing of (future) sentient AIs, particularly given the uncontroversial nature of sentience as a criterion. However, people are much more skeptical of AIs capacity for sentience than their capacity for cognition. For example, Pauketat and Anthis [59] found that, on a scale from 0 (not at all*)* to 100 (very much), the mean response to whether future AIs can have emotions was 38.6 (standard deviation = 30.4), compared to a mean of 70.9 (standard deviation = 22.7) for cognition.[41] It may be very difficult to show with enough confidence that certain AIs are sentient, if we can even have such confidence ourselves. Thus, an excessive focus on sentience may be too high or uncertain of a bar for AI systems to meet. However, it is plausible that AIs will convincingly develop capacities associated with cognition before clearly demonstrating their sentience. Since cognition is widely considered to be a sufficient criterion for moral standing anyway, it may make sense to increase focus on this criterion so that AIs generally are more likely to be accepted as the types of entities that can have moral standing. Once some AIs are granted moral standing, it may be easier for society to go on to extend the possibility of sentience in AIs and moral standing to AIs on this basis.[42]

The second consideration concerns future non-sentient AIs that are highly cognitively sophisticated, at a level roughly equal to present-day humans. We may not think that such entities have the required capacities for moral standing (and they may not think that we do). However, if we interact with such entities in a shared social environment, it will likely be important that we cooperate with them in a similar way to how we try to cooperate with other humans today. This will entail coming up with arrangements in which we reciprocally take each other's interests into account in our decision making. In other words, it will entail treating each other as if we have moral standing.

The third consideration is about non-sentient AIs that are more powerful than humans. Such a scenario could arise if, for example, AIs become more intelligent, run at much faster processing speeds, or exist in far greater numbers than humans. Whether we grant moral standing to these AIs is likely to be of little relevance from their perspective; it would arguably be analogous to chimpanzees extending moral standing to humans today. Of more relevance is whether the AIs would grant moral standing to humans and other sentient entities. Bostrom and Shulman [7] suggest that our relations with the precursors of such advanced AIs (perhaps such as those described in the previous paragraph) may be very important from a practical and prudential perspective, and that a cooperative scheme with "high-potential early AIs" is likely to be more positive than an uncooperative one. This may be because a cooperative scheme with early AIs may lead to values and institutions favorable to humans and other sentient entities that persist in the long run, even when they are no longer necessary for the later-stage more powerful AIs. As Bostrom and Shulman [7] note, even if a tiny fraction of total resources available to future extremely powerful AI systems are allocated to humans, this could result in a very positive outcomes from a human-centric perspective. Thus, having more expansive criteria for moral standing could be in our own long run interest.

As with the previous section on uncertainty, the discussion here is far from conclusive. However, I think that these preliminary considerations provide further support to the view that we should grant moral standing to at least some non-sentient AIs.

[40] To be more concrete, assume we have 70% credence in view (1) and 30% credence in view (2). Now imagine that the value of damaging a non-sentient AI is + 1 on view (1) and − 10 on view (2), and that the value of not damaging the non-sentient AI is − 1 on view (1) and 0 on view (2). The expected value of choosing to damage the non-sentient AI is (1 * 70%) + (− 10 * 30%) = − 2.3 and the expected value of choosing not to damage the non-sentient AI is (− 1 * 70%) + (0 * 30%) = − 0.7. Since the expected value of damaging the AI is lower than not doing so, we should avoid doing so.

[41] See Appendix A of Pauketat and Anthis [59].

[42] This consideration will of course be stronger if we have higher confidence in the view that non-sentient AIs can have moral standing.

7 Implications

In contrast to some recent studies, I argue against the view that sentience [26] or even consciousness [52] is necessary for an AI to qualify for moral standing. Instead, I suggest that sentience is sufficient, and that some non-sentient and some non-conscious AIs may also qualify for moral standing. This has the implication that the issue of AI moral standing may be more important, in terms of its scale and urgency, than on views that take either sentience or consciousness to be necessary. It may be greater in scale because the view in the present article plausibly allows more AIs to qualify for moral standing than if either sentience or consciousness is necessary. It may be more urgent because non-conscious cognitively complex AIs that are candidates for moral standing will plausibly be created sooner than sentient or conscious AIs. This seems plausible because while researchers are currently able to build AIs with a range of cognitive capacities, there is limited understanding of how to implement sentience or consciousness in them [49].

A second implication concerns the growing literature on designing and promoting social policies that take into account the interests of sentient AIs (e.g., [51, 80]. According to the arguments in this article, while this literature addresses an important topic—sentience is sufficient for AIs to qualify for moral standing—it may be underinclusive of non-sentient AIs that also qualify for moral standing. Researchers in this field should consider broadening their scope to design and promote policies that consider all AIs with morally relevant interests, not only sentient AIs.

A third implication concerns the argument that AIs cannot be sentient or conscious and so cannot qualify for moral standing (e.g., [39, 42, 73]). The arguments in the present article suggest that even if it is impossible for AIs to be sentient or conscious, some of them might still qualify for moral standing. This means that even people who are skeptical of the possibility of AI sentience or consciousness should take the question of AI moral standing seriously.

8 Limitations and future directions

This article has several important limitations. First, it is worth highlighting two of the assumptions made. While not strictly entailed by my definition of moral standing, I assume that an entity must have interests to qualify for moral standing. In other words, I assume welfarism [44]. This excludes the possibility of some entities, such as a painting, a mountain, or an AI without interests, from qualifying for moral standing. While I think this is a defensible position, people who think such entities can qualify for moral standing could take the meta-criterion proposed in Sect. 3 as only a sufficient condition, and propose separate criteria for entities without interests. I also assume naturalism, in the sense that I only consider criteria that would be accepted in a modern scientific framework. This excludes criteria such as "having a soul." As with the previous assumption, I think this is defensible, but people who hold alternative views may wish to consider such criteria in future research.

Second, there are several uncertainties with the arguments made in this article. The first uncertainty concerns the thought experiments used to support the view that some non-sentient AIs should be granted moral standing. As noted in Sects. 2.2 and 2.3, some of these thought experiments make it difficult to isolate the effects of non-sentient capacities on our moral judgments. For example, Kagan's [43] non-sentient robots behave like they are sentient, making it possible that our judgment that they should have moral standing is based on the precaution that they might in fact be sentient. Future research should devise thought experiments that better isolate the effects of non-sentient capacities on our moral judgments. There is also uncertainty about the implications of different ways of dealing with uncertainty about which capacities should qualify an entity for moral standing, discussed in Sect. 5, and about the strategic considerations of granting moral standing to AIs, discussed in Sect. 6. As noted in those sections, while I think both of these provide some support for granting moral standing to some non-sentient AIs, the considerations are preliminary. Future research should conduct more in-depth analysis of these issues.

A third limitation is that the meta-criterion I argue for in Sect. 3, that an entity should qualify for moral standing if it has interests that "matter to" it, is imprecise.[43] What exactly does it mean for interests to "matter to" an entity, and how can we judge whether an entity has such interests? While I think the meta-criterion is sufficiently precise to allow us to make useful heuristic judgments about which entities satisfy it, as I do in Sect. 3, an important topic of future research is to develop a more systematic method for making judgments about which entities can be said to have interests that "matter to" them.

A fourth limitation is that the notion of moral standing used in this article is itself limited. As defined in the Introduction, having moral standing requires that an entity's interests are taken into account "to at least some degree." But this is consistent with an entity being granted minimal moral consideration, or with its interests always being overridden by the interests of other entities. Therefore, future research should go beyond the question of moral standing

[43] While this is a limitation of the proposed meta-criterion, it arguably also applies to each of the individual criteria as well. For example, see Anthis [1] for the imprecision of "consciousness" and "sentience" as criteria.

to address how much the interests of different AIs matter, and how to appropriately take those interests into account in our decision making. Given the potential for future AIs to develop vastly more advanced morally relevant capacities than present-day biological entities [65, 68] these questions are particularly challenging and important.

9 Conclusion

In conclusion, while I think sentience is sufficient for moral standing, I think that there is a strong case for thinking that some non-sentient AIs, such as those that are conscious and have non-valenced preferences and goals, and those that are non-conscious but have sufficiently cognitively complex preferences and goals, should also be granted moral standing. Taking into account uncertainty and strategic considerations may further support for doing so. If correct, this has the implications that the issue of AI moral standing may be more important in terms of its scale and urgency than if sentience or consciousness is necessary, that researchers working on designing policies inclusive of sentient AIs should broaden their scope to include all AIs with morally relevant interests, and that even those who think AIs cannot be sentient or conscious should take the issue of AI moral standing seriously. However, much uncertainty remains, making this an important topic of future research.

Acknowledgements Many thanks to Reuben Adams, Jacy Reese Anthis, Patrick Butlin, Max Carpendale, Mark Coeckelbergh, Michael Dello-Iacovo, Oscar Horta, Katerina Manoli, Clotilde de Maricourt, Eze Paez, Janet Pauketat, Bradford Saad, and Jason Schukraft for helpful comments and discussion.

Author contributions The author confirms being the sole contributor of this article.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability Not applicable.

Declarations

Conflict of interest The author has no competing interests to declare.

Ethical approval Not applicable.

Consent for publication Not applicable.

References

1. Anthis, J.R.: Consciousness semanticism: a precise eliminativist theory of consciousness. In: Klimov, V.V., Kelley, D.J. (eds.) Biologically inspired cognitive architectures 2021, pp. 20–41. Springer International Publishing (2022). https://doi.org/10.1007/978-3-030-96993-6_3
2. Baertschi, B.: The moral status of artificial life. Environ. Values **21**(1), 5–18 (2012). https://doi.org/10.3197/096327112X13225063227907
3. Basl, J.: Machines as moral patients we shouldn't care about (yet): the interests and welfare of current machines. Philos. Technol. **27**(1), 79–96 (2014). https://doi.org/10.1007/s13347-013-0122-y
4. Birch, J.: Animal sentience and the precautionary principle. Anim. Sentience (2017). https://doi.org/10.51291/2377-7478.1200
5. Blackmore, S., Troscianko, E.T.: Consciousness: an introduction, 3rd edn. Routledge (2018). https://doi.org/10.4324/9781315755021
6. Block, N.: On a confusion about a function of consciousness. Behav. Brain Sci. **18**(2), 227–247 (1995). https://doi.org/10.1017/S0140525X00038188
7. Bostrom, N., Shulman, C.: Propositions concerning digital minds and society. (2022). https://nickbostrom.com/propositions.pdf. Accessed 25 Sept 2022
8. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. In: The cambridge handbook of artificial intelligence, pp. 316–334. Cambridge University Press (2014)
9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Amodei, D.: Language Models are Few-Shot Learners (2020). (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165
10. Buchanan, A.: Moral status and human enhancement. Philos. Public Aff. **37**(4), 346–381 (2009)
11. Chalmers, D.J.: Reality+: virtual worlds and the problems of philosophy. Penguin UK (2022)
12. Coeckelbergh, M.: Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf. Technol. **12**(3), 209–221 (2010). https://doi.org/10.1007/s10676-010-9235-5
13. Coeckelbergh, M.: The moral standing of machines: towards a relational and non-Cartesian moral hermeneutics. Philos. Technol. **27**(1), 61–77 (2014). https://doi.org/10.1007/s13347-013-0133-8
14. Coeckelbergh, M.: Should we treat teddy bear 2.0 as a Kantian dog? Four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. Minds Mach. **31**(3), 337–360 (2021). https://doi.org/10.1007/s11023-020-09554-3
15. Cotton-Barratt, O., Greaves, H.: A bargaining-theoretic approach to moral uncertainty. Global Priorities Institute. (2019). https://globalprioritiesinstitute.org/a-bargaining-theoretic-approach-to-moral-uncertainty/. Accessed 25 Sept 2022
16. Crisp, R.: Well-being. In E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Winter 2021). Metaphysics Research Lab, Stanford University. (2021). https://plato.stanford.edu/archives/win2021/entries/well-being/. Accessed 23 Sept 2022

17. Cudd, A., Eftekhari, S.: Contractarianism. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2021). Metaphysics Research Lab, Stanford University. (2021). https://plato.stanford.edu/archives/win2021/entries/contractarianism/. Accessed 23 Sept 2022
18. Danaher, J.: Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci. Eng. Ethics **26**(4), 2023–2049 (2020). https://doi.org/10.1007/s11948-019-00119-x
19. Danaher, J.: What matters for moral status: behavioral or cognitive equivalence? Camb. Q. Healthc. Ethics **30**(3), 472–478 (2021). https://doi.org/10.1017/S0963180120001024
20. DeGrazia, D.: Taking animals seriously: mental life and moral status. Cambridge University Press (1996). https://doi.org/10.1017/CBO9781139172967
21. DeGrazia, D.: Great apes, dolphins, and the concept of personhood. South. J. Philos. **35**(3), 301–320 (1997). https://doi.org/10.1111/j.2041-6962.1997.tb00839.x
22. DeGrazia, D.: Robots with moral status? Perspect. Biol. Med. **65**(1), 73–88 (2022). https://doi.org/10.1353/pbm.2022.0004
23. Degrazia, D., Millum, J. (eds.): Moral status. A theory of bioethics, pp. 175–213. Cambridge University Press (2021). https://doi.org/10.1017/9781009026710.007
24. Floridi, L.: Information ethics: On the philosophical foundation of computer ethics. Ethics Inf. Technol. **1**(1), 33–52 (1999). https://doi.org/10.1023/A:1010018611096
25. Francione, G.L., Charlton, A.: Animal rights: the abolitionist approach. Exempla Press (2015)
26. Gibert, M., Martin, D.: In search of the moral status of AI: why sentience is a strong argument. AI Soc. **37**(1), 319–330 (2022). https://doi.org/10.1007/s00146-021-01179-z
27. Godfrey-Smith, P.: Varieties of subjectivity. Philos. Sci. **87**(5), 1150–1159 (2020). https://doi.org/10.1086/710541
28. Goff, P., Seager, W., Allen-Hermanson, S.: Panpsychism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. (2022). https://plato.stanford.edu/archives/sum2022/entries/panpsychism/
29. Goodpaster, K.E.: On being morally considerable. J. Philos. **75**(6), 308–325 (1978). https://doi.org/10.2307/2025709
30. Gordon, J.-S., Pasvenskiene, A.: Human rights for robots? A literature review. AI Ethics **1**(4), 579–591 (2021). https://doi.org/10.1007/s43681-021-00050-7
31. Graziano, M.S.A.: Rethinking consciousness: a scientific theory of subjective experience. W. W. Norton & Company (2019)
32. Gruen, L.: Conscious animals and the value of experience. In: Gardiner, S.M., Thompson, A. (eds.) The Oxford handbook of environmental ethics. Oxford University Press (2017). https://doi.org/10.1093/oxfordhb/9780199941339.013.9
33. Gunkel, D.J.: The other question: can and should robots have rights? Ethics Inf. Technol. **20**(2), 87–99 (2018). https://doi.org/10.1007/s10676-017-9442-4
34. Gustafsson, J.E., Torpman, O.: In defence of my favourite theory. Pac. Philos. Q. **95**(2), 159–174 (2014). https://doi.org/10.1111/papq.12022
35. Harman, E.: The potentiality problem. Philos. Stud. **114**(1/2), 173–198 (2003)
36. Harris, J., Anthis, J.R.: The moral consideration of artificial entities: a literature review. Sci. Eng. Ethics **27**(4), 53 (2021). https://doi.org/10.1007/s11948-021-00331-8
37. Horta, O.: The scope of the argument from species overlap. J. Appl. Philos. **31**(2), 142–154 (2014). https://doi.org/10.1111/japp.12051
38. Horta, O.: Moral considerability and the argument from relevance. J. Agric. Environ. Ethics **31**(3), 369–388 (2018). https://doi.org/10.1007/s10806-018-9730-y
39. Hsing, D. Artificial consciousness is impossible. Towards Data Science. (2021). https://towardsdatascience.com/artificial-consciousness-is-impossible-c1b2ab0bdc46
40. Jaworska, A.: Caring and full moral standing. Ethics **117**(3), 460–497 (2007). https://doi.org/10.1086/512780
41. Jaworska, A., Tannenbaum, J.: The Grounds of Moral Status. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2021). Metaphysics Research Lab, Stanford University. (2021). https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/
42. Johnson, D.G., Verdicchio, M.: Why robots should not be treated like animals. Ethics Inf. Technol. **20**(4), 291–301 (2018). https://doi.org/10.1007/s10676-018-9481-5
43. Kagan, S.: How to count animals, more or less. Oxford University Press (2019)
44. Keller, S.: Welfarism. Philos. Compass **4**(1), 82–95 (2009). https://doi.org/10.1111/j.1747-9991.2008.00196.x
45. Kirk, R.: Zombies. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2021). Metaphysics Research Lab, Stanford University. (2021). https://plato.stanford.edu/archives/spr2021/entries/zombies/
46. Knott, A., Sagar, M., Takac, M.: The ethics of interaction with neurorobotic agents: a case study with BabyX. AI and Ethics **2**(1), 115–128 (2022). https://doi.org/10.1007/s43681-021-00076-x
47. Korsgaard, C.: Fellow creatures: kantian ethics and our duties to animals. (2004). https://dash.harvard.edu/handle/1/3198692
48. MacAskill, M., Bykvist, K., Ord, T.: Moral uncertainty. Oxford University Press (2020). https://doi.org/10.1093/oso/9780198722274.001.0001
49. McDermott, D.: Artificial intelligence and consciousness. In: Zelazo, P.D., Moscovitch, M., Thompson, E. (eds.) Cambridge handbook of consciousness, pp. 117–150. Cambridge University Press, Cambridge (2007)
50. McMahan, J.: The ethics of killing: problems at the margins of life. Oxford University Press (2002)
51. Metzinger, T.: Artificial suffering: an argument for a global moratorium on synthetic phenomenology. J. Artif. Intell. Conscious. **08**(01), 43–66 (2021). https://doi.org/10.1142/S270507852150003X
52. Mosakas, K.: On the moral status of social robots: considering the consciousness criterion. AI Soc. **36**(2), 429–443 (2021). https://doi.org/10.1007/s00146-020-01002-1
53. Neely, E.L.: Machines and the moral community. Philos. Technol. **27**(1), 97–111 (2014). https://doi.org/10.1007/s13347-013-0114-y
54. Newberry, T., Ord, T.: The parliamentary approach to moral uncertainty. Future of Humanity Institute, Technical Report #2021–2 (2021)
55. Noddings, N.: Caring: a relational approach to ethics and moral education, 2nd edn. University of California Press (2013)
56. Nussbaum, M.C.: The moral status of animals. Chron. High. Educ. **52**(22), B6-8 (2006)
57. Parfit, D.: Reasons and persons. OUP Oxford (1984)
58. Pauketat, J. V.: The terminology of artificial sentience. PsyArXiv. (2021). https://doi.org/10.31234/osf.io/sujwf
59. Pauketat, J.V.T., Anthis, J.R.: Predicting the moral consideration of artificial intelligences. Comput Hum Behav **136**, 107372 (2022). https://doi.org/10.1016/j.chb.2022.107372
60. Peterson, M.: An introduction to decision theory. Higher Education from Cambridge University Press; Cambridge University Press (2017). https://doi.org/10.1017/9781316585061
61. Regan, T.: The case for animal rights. University of California Press (2004)

62. Rodogno, R.: Sentientism, wellbeing, and environmentalism. J. Appl. Philos. **27**(1), 84–99 (2010). https://doi.org/10.1111/j.1468-5930.2009.00475.x
63. Roelofs, L.: Sentientism, motivation, and philosophical Vulcans. Pac. Philos. Q. (2022). https://doi.org/10.1111/papq.12420
64. Scherer, D.: Anthropocentrism, atomism, and environmental ethics. Environ. Ethics **4**(2), 115–123 (1982). https://doi.org/10.5840/enviroethics19824220
65. Schwitzgebel, E., Garza, M.: A defense of the rights of artificial intelligences. Midwest Stud. Philos. **39**, 98–119 (2015). https://doi.org/10.1111/misp.12032
66. Sebo, J.: The moral problem of other minds. Harv. Rev. Philos. **25**, 51–70 (2018). https://doi.org/10.5840/harvardreview20185913
67. Shevlin, H.: How could we know when a robot was a moral patient? Camb. Q. Healthc. Ethics **30**(3), 459–471 (2021). https://doi.org/10.1017/S0963180120001012
68. Shulman, C., Bostrom, N.: Sharing the world with digital minds. In: Clarke, S., Zohny, H., Savulescu, J. (eds.) Rethinking moral status, pp. 306–326. Oxford University Press (2021). https://doi.org/10.1093/oso/9780192894076.003.0018
69. Singer, P.: Practical ethics. Higher Education from Cambridge University Press; Cambridge University Press (2011). https://doi.org/10.1017/CBO9780511975950
70. Singer, P., Sagan, A.: When robots have feelings. The Guardian. (2009). https://www.theguardian.com/commentisfree/2009/dec/14/rage-against-machines-robots. Accessed 3 Sept 2022
71. Sinnott-Armstrong, W., Conitzer, V.: How much moral status could artificial intelligence ever achieve? In: Clarke, S., Zohny, H., Savulescu, J. (eds.) Rethinking moral status, pp. 269–289. Oxford University Press (2021). https://doi.org/10.1093/oso/9780192894076.003.0016
72. Smids, J.: Danaher's ethical behaviourism: an adequate guide to assessing the moral status of a robot? Sci. Eng. Ethics **26**(5), 2849–2866 (2020). https://doi.org/10.1007/s11948-020-00230-4
73. Sparrow, R.: The turing triage test. Ethics Inf. Technol. **6**(4), 203–213 (2004). https://doi.org/10.1007/s10676-004-6491-2
74. Stone, J.: Why potentiality matters. Can. J. Philos. **17**(4), 815–829 (1987). https://doi.org/10.1080/00455091.1987.10715920
75. Torrance, S.: Artificial consciousness and artificial ethics: between realism and social relationism. In: Machine ethics and robot ethics. Routledge (2017)
76. Wang, X., Krumhuber, E.G.: Mind perception of robots varies with their economic versus social function. Front. Psychol. (2018). https://doi.org/10.3389/fpsyg.2018.01230
77. Warren, M.A.: Difficulties with the strong animal rights position. Between Species **2**(4), 4 (1986)
78. Warren, M.A.: Moral status: obligations to persons and other living things. Clarendon Press (1997)
79. Waytz, A., Norton, M.I.: Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs. Emotion **14**, 434–444 (2014). https://doi.org/10.1037/a0036054
80. Ziesche, S., Yampolskiy, R.: Towards AI welfare science and policies. Big Data Cogn. Comput. (2019). https://doi.org/10.3390/bdcc3010002