ROBONOMICS
The Journal of the Automated Economy
https://journal.robonomics.science

# The Full Rights Dilemma for AI Systems of Debatable Moral Personhood

Eric Schwitzgebel [1] ✉

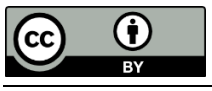
**Abstract**
An Artificially Intelligent system (an AI) has debatable moral personhood if it is epistemically possible either that the AI is a moral person or that it falls far short of personhood. Debatable moral personhood is a likely outcome of AI development and might arise soon. Debatable AI personhood throws us into a catastrophic moral dilemma: Either treat the systems as moral persons and risk sacrificing real human interests for the sake of entities without interests worth the sacrifice, or do not treat the systems as moral persons and risk perpetrating grievous moral wrongs against them. The moral issues become even more perplexing if we consider cases of possibly conscious AI that are subhuman, superhuman, or highly divergent from us in their morally relevant properties.

**Keywords:** artificial intelligence, ethics, persons, robot rights, transhumanism

**Type:** Article

**Citation:** Schwitzgebel, E. (2023). The Full Rights Dilemma for AI Systems of Debatable Moral Personhood. *ROBONOMICS: The Journal of the Automated Economy,* 4, 32

[1] Department of Philosophy, University of California, Riverside, Riverside, CA 92521, USA; email: eschwitz@ucr.edu
✉ Corresponding author



**Introduction**

We might soon build artificially intelligent entities – AIs – of debatable moral personhood. We will then need to decide whether to grant these entities the full range of rights[1] and moral consideration that we normally grant to fellow humans. Our systems and habits of ethical thinking are currently as unprepared for this decision as medieval physics was for space flight.

If there is even a small chance that some technological leap could soon produce AI systems with a reasonable claim to personhood, the issue deserves careful consideration in advance. We will have ushered a new type of entity into existence – an entity *perhaps* as morally significant as *Homo sapiens,* and one likely to possess radically new forms of existence. Few human achievements have such potential moral importance and such potential for moral catastrophe.

An entity has *debatable moral personhood,* as I intend the phrase, if it is reasonable to think that the entity might be a person in the sense of deserving the same type of moral consideration that we normally give, or ought to give, to human beings, and if it is also reasonable to think that the entity might fall far short of deserving such moral consideration. I intend "personhood" as a rich, demanding moral concept.[2] If an entity is a moral person, they normally deserve to be treated as an equal of other persons, including for example – to the extent appropriate to their situation and capacities – deserving "human" rights, care and concern similar to that of other people, and equal protection under the law. Personhood, in this sense, entails moral status, moral standing, or moral considerability fully equal to that of ordinary human beings (Jaworska and Tannenbaum 2013/2021). By "moral personhood" I do not, for example, mean merely the legal personhood sometimes attributed to corporations for certain purposes.[3]

An AI's personhood is "debatable", as I will use the term, if it is reasonable to think that the AI might be a person but also reasonable to think that the AI might fall far short of personhood. Substantial doubt is appropriate – not just minor doubts about the precise place to draw the line in a borderline case. Note that debatable personhood in this sense is both epistemic and relational: An entity's status as a person is debatable if *we* (we in some epistemic community, however defined) are not compelled, given our available epistemic resources, either to reject its personhood or to reject the possibility that it falls far short.[4] Other entities or communities, or our future selves, with different epistemic resources, might know perfectly well whether the entity is a person. Debatable personhood is thus not an intrinsic feature of an entity but rather a feature of our epistemic relationship to that entity.

I will defend four theses. First, debatable personhood is a likely outcome of AI development. Second, AI systems of debatable personhood might arise soon. Third, debatable AI personhood throws us into a catastrophic moral dilemma: Either treat the systems as moral persons and risk sacrificing real human interests for the sake of entities without interests worth the sacrifice, or don't treat the systems as moral persons and risk perpetrating grievous moral wrongs against them. Fourth, the moral issues become even more perplexing if we consider cases of possibly conscious AI that are subhuman, superhuman, or highly divergent from us in their morally relevant properties.

**1. Non-Persons: The Near Future of Humanlike AI**

Near-future cases will be non-persons. Consider GPT-3, ChatGPT, and GPT-4 – computer programs that can produce strikingly realistic linguistic outputs after receiving linguistic inputs.[5] Ask a language model of this sort to write a poem and it will write a poem. Ask it to play chess and it will produce a series of plausible chess moves. Feed it the title of a story and the byline of a famous author – for example, "The Importance of Being on Twitter by Jerome K. Jerome" – and it will produce clever prose in that author's style (Klingemann, 2020):

> *The Importance of Being on Twitter*
> by Jerome K. Jerome
> London, Summer 1897
>
> It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage.

Models of this sort achieve all of this without being specifically trained on tasks of this sort, though for the best results human users will typically choose the best among a handful of outputs. A group of philosophers wrote opinion pieces about the significance of GPT-3 and then fed it those pieces as input. It produced an intelligent-seeming, substantive reply, including passages like (Weinberg, 2020):

> To be clear, I am not a person. I am not self-aware. I am not conscious. I can't feel pain. I don't enjoy anything. I am a cold, calculating machine designed to simulate human response and to predict the probability of certain outcomes. The only reason I am responding is to defend my honor.

It almost seems to have a sense of humor.

Now imagine a GPT-3 mall cop. Actually, assume a few more generations of technological improvement – GPT-6 maybe. Give it speech-to-text and text-to-speech so that it can respond to and produce auditory language. Mount it on a small autonomous vehicle, like a delivery bot, but with a humanoid form. Give it camera eyes and visual object recognition as context for its speech outputs. To keep it friendly, inquisitive, and not too weird, give it some behavioral constraints and additional training on a database of mall-like interactions, plus a good, updatable map of the mall and instructions not to leave the area. Give it a socially interactive face, like MIT's "Kismet" robot.[6] Give it some short-term and long-term memory. Finally, give it responsiveness to tactile inputs, a map of its bodily boundaries, and hands with five-finger grasping. All of this is technologically feasible now, though expensive. Such a robot could be built within a few years.

This robot will of course chat with the mall patrons. It will comment politely on their purchases, tell jokes, complain about the weather, and give them directions if they are lost. Some patrons will avoid interaction, but others – like my daughter at age eight when she discovered the "Siri" chatbot on my iPhone – will enjoy interacting with it. They will ask what it's like to be a mall cop, and it will say something sensible in reply. They will ask what it does on vacation, and it might tell amusing lies about Tahiti or tales of sleeping in the mall basement. They will ask whether it likes this shirt or this other one, and then they'll buy the shirt it prefers. They will ask if it's conscious and if it has feelings and is a person just like them, and it might say no or it might say yes.

Such a robot could reconnect with previous conversation partners. Using facial recognition software, it might recognize patrons' faces. It might then retrieve stored records of previous conversations with that patron. Based on word valences and its reading of emotional facial expressions, it might assess patrons' openness to further conversation. Using previous conversations as a context for new speech, it might roll or stride forward with "Hi, Natalie! Good to see you again. I hope you're enjoying that shirt you bought last Wednesday!" Based on facial and linguistic cues that suggest that the patron is reacting positively or negatively, it could modify its reactions on the fly and further tune future reactions, both to that person in particular and to mall patrons in general. It could react appropriately to hostility. A blow to the chest might trigger a fear face, withdrawal, and a plaintive plea to be left alone. It might cower and flee quite convincingly and pathetically, wailing and calling desperately for its friends to help. (Assume this robot is not designed to defend itself with physical aggression.)

Maybe our mall patroller falls to its knees in front of a sympathetic bystander, begging for protection against a crowbar-wielding Luddite.

If the robot speaks well enough and looks human enough, some people will presumably think that it has genuine feelings and experiences – "phenomenal consciousness" in the philosopher's sense.[7] They will think it is sentient, that is, capable of feeling pleasure and pain. If the robot is threatened or abused, some people will be emotionally moved by its plight – not merely as we can be moved by the plight of a character in a novel or video game, and not merely as we can be disgusted by the callous destruction of valuable property. Some people will believe that the robot is genuinely suffering under abuse, or genuinely happy to see a friend again, genuinely sad to hear that an acquaintance has died, genuinely surprised and angry when a vandal breaks a shop window.

Many of these same people will presumably also think that the robot should not be treated in certain ways. If they think it is genuinely capable of suffering, they will probably also think that we ought not needlessly make it suffer. They will think the robot has at least some limited rights, some intrinsic moral standing. They will think that it is not merely property that its owner should feel free to abuse or destroy at will without good reason. This isn't necessarily yet full moral personhood. For example, most people think that dogs have a limited moral standing that is short of moral personhood. Early advocates of AI rights might argue for less-than-person status similar to that of non-human animals or of a *sui generis* sort (E.g. Basl, 2013, 2014; Gunkel, 2018).

Now *you* might think it's clear that near-future robots, constructed this way, could not really suffer, could not possibly have genuine sentience or humanlike consciousness of the sort that warrants reactions of sympathetic concern for its pain. Philosophers, psychologists, computer programmers, neuroscientists, and experts on consciousness might be near consensus that a robot designed as I have just described would be no more conscious than a desktop computer. Leading experts might know that it just mixes a few technologies we already possess. There might be no prominent theory of consciousness that awards the machine high marks.

But not everyone will agree with skeptical experts – perhaps especially among the younger generation. Recent survey results, for example, suggest that the large majority of U.S. and Canadian respondents under age 30 think that robots may someday really experience pleasure and pain – a much less common view among older respondents (De Graaf, Hindricks, & Hindricks, 2021). Studies by Kate Darling suggest that ordinary research participants are already reluctant to smash little robot bugs after those bugs have been given names that personify them (Darling, 2017; see also Darling, 2016, 2021). Imagine how much more reluctant people (most people) might be if the robot is not a mere bug but something with humanoid form, an emotionally expressive face, and humanlike speech, pleading for its life.

Soldiers already grow attached to battlefield robots, burying them, "promoting" them, sometimes even risking their lives for them (Gerreau, 2007; Singer, 2009; Carpenter, 2016; Gunkel, 2018). People fall in love with, or appear to fall in love with – or at least become seriously emotionally attached to – currently existing chatbots like Replika (Shevlin, 2021). There already is a "Robot Rights" movement. There already is a society modeled on the famous animal rights organization PETA (People for the Ethical Treatment of Animals), called People for the Ethical Treatment of Reinforcement Learners. These are currently small movements. As AI becomes more sophisticated, and as chatbots start sounding more and more human, these movements will presumably gain more adherents, especially among people with liberal views of AI consciousness or sentience.

If AI technology continues to improve, eventually robot rights activists will form a large enough group to influence corporate or government policy. They might demand that malls treat their robot patrollers in certain ways. They might insist that companion robots for children and the elderly be protected from certain kinds of cruelty and abuse. They might insist that care-and-use committees evaluate the ethics of research on robots in the same way that such committees currently evaluate research on non-human vertebrates.[8] If the machines

become human enough in their outward behavior, some people will treat them as friends, fall in love, liberate them from servitude, and eventually demand even that robots be given full legal personhood and morally respected to the same high degree we ought ordinarily to respect fellow humans. That is, they will see these robots as moral persons. This might happen even while large groups of our fellow humans remain morally devalued or neglected.

Even if current AI systems and our possible near-future GPT-6 mall patroller deserve some amount of moral consideration, they presumably lack even debatable personhood. They are not moral persons. Well-informed people will, I assume, be epistemically compelled to regard such systems as far short of genuinely deserving full humanlike rights or moral consideration as our equals. However, if technology continues to improve, eventually it will become reasonable to wonder whether some of our AI systems might really be persons. As soon as that happens, those AI systems will possess debatable personhood.[9]

**2. Two Ethical Assumptions**

I will now make two ethical assumptions which I hope the reader will find plausible. The first is that it is not in principle impossible to build an AI system who is a person in the intended moral sense of that term. The second is that the presence or absence of consciousness – phenomenal consciousness, in the sense of the term central to philosophy and consciousness studies[10] – is crucial in the assessment of whether an AI system is a moral person in the sense of deserving moral consideration similar to that of ordinary human beings.

In other work in collaboration with Mara Garza, I have defended the first assumption at length (Schwitzgebel & Garza, 2015). Our core argument is as follows.

*Premise 1*: If Entity A deserves some particular degree of moral consideration and Entity B does not deserve that same degree of moral consideration, there must be some *relevant difference* between the two entities that grounds this difference in moral status.
*Premise 2*: There are possible AIs who do not differ in any such relevant respects from human beings.
*Conclusion*: Therefore, there are possible AIs who deserve a degree of moral consideration similar to that of human beings.

The conclusion follows logically from the premises, and Premise 1 seems hard to deny. So if there is a weakness in this argument, it is probably Premise 2. I have heard four main objections to the idea expressed in that premise: (1.) that any AI would necessarily lack some crucial feature such as consciousness, freedom, or creativity; (2.) that any AI would necessarily be outside of our central circle of concern because it does not belong to our species; (3.) that AI would lack personhood because it can be duplicated; and (4.) that AI would have reduced moral claims on us because it owes its existence to us.

None of these objections survive scrutiny. Against the first objection, such in-principle AI skepticism (unless, perhaps, grounded in theistic assumptions about the necessity of God's hand in creating consciousness) seems to disregard the possibly wide diversity of future technological approaches, including possible forms of artificial life. Even John Searle and Roger Penrose, perhaps the most famous AI skeptics, allow that some future AI systems (designed very differently from 20th century computers) might have as much consciousness, freedom, and creativity as we do.[11] The species-based second objection constitutes noxious bigotry that would unjustly devalue AI friends and family who (if the response to the first objection stands) might be no different from us in any relevant psychological or social characteristics and might consequently be fully integrated into our society.[12] The duplicability-based third objection falsely assumes that AI must be duplicable rather than relying on fragile or uncontrollable processes, and it overrates the value of non-duplicability. Finally, the objection from existential debt is exactly backwards: If we create genuinely humanlike AI, socially and psychologically similar to us, we will owe *more* to it than we owe to human strangers, since we will have been responsible for its existence and

presumably also to a substantial extent for its happy or miserable state – a relationship comparable to that between parents and children.[13]

My second assumption, concerning the importance of consciousness, divides into two sub-claims:

*Claim A*: Any AI system that entirely lacks conscious experiences is far short of being a person.
*Claim B*: Any AI system with a fully humanlike range of conscious capacities and experiences is a person.

The question of what grounds moral standing or moral personhood is huge and fraught. Simplifying, approaches divide into two broad camps between which it is reasonable to remain undecided. Utilitarian views, historically grounded in the work of Jeremy Bentham and John Stuart Mill, hold that what matters is an entity's capacity for pleasure and suffering. Anything capable of sufficient pleasure and suffering deserves humanlike moral consideration.[14] Other views hold that instead what matters is the capacity for a certain type of rational thought, or other types of "higher" cognitive capacities, or creative or social flourishing. Or rather, to speak more carefully, since most philosophers regard infants and people with severe cognitive disabilities as deserving of no less moral regard than ordinary adults, what is necessary on this view is something like the right kind of potentiality for such cognition or flourishing, whether future, past, counterfactual, or by possession of the right essence or group membership (Sussman, 2003; Jaworska & Tannenbaum, 2013/2021; Korsgaard, 2018; Kagan, 2019; Floris, 2021).

Philosophical views about the grounds of moral standing sometimes do not explicitly specify that the pleasure and suffering, the rational cognition, or the human flourishing must belong to an entity that is phenomenally conscious in the standard philosophical and "consciousness studies" sense of the term "conscious". However, most theorists would accept consciousness (or at least the potentiality for it) as a necessary condition of full personhood.[15] The second assumption of this section is that, regardless of the details about the specific bases of moral standing (utilitarian or otherwise), the presence or absence of consciousness is extremely relevant in that a system that utterly lacks consciousness is far short of moral personhood (Claim A) and a system that has a fully humanlike range of conscious capacities and experiences is a person (Claim B). On this issue, many philosophers from diverse camps will agree.

Imagine an AI system that is entirely nonconscious but otherwise as similar as possible to an ordinary adult human being. It might be superficially human and at least roughly humanlike in its outward behavior – like the mall patroller, further updated – but suppose, for the sake of argument, that we know it completely lacks any capacity for consciousness. It never has any conscious experiences of pleasure or pain, never has any conscious thoughts or imagery, never forms a conscious plan, never consciously thinks anything through, never has any visual experiences or auditory experiences, no sensations of hunger, no feelings of comfort or discomfort, no experiences of alarm or compassion – no conscious experiences at all, ever. Such an AI might be amazing! It might be a truly fantastic piece of machinery, worth valuing and preserving on those grounds. But it would not, I am assuming, be a *person* in the full moral sense of the term. That is Claim A.[16]

Claim B complements Claim A. Imagine an AI very differently constructed from an ordinary human being, yet having the full range of conscious experiences that human beings enjoy. To keep the thought experiment neutral among simple utilitarian approaches and approaches to moral standing that require specific, complex, human-like, cognitive capacities, imagine that this AI system has not only the capacity for immense pleasure and suffering but also all of the complex, human-like, cognitive capacities that are plausibly relevant on any leading theory. In other words, imagine this AI system, despite perhaps having a radically different internal constitution and outward form, to be experientially very like us. More specifically: It is capable of humanlike pleasure at success and suffering at loss. When injured, it feels pain as sharply as we do. It has visual and auditory consciousness of its environment, which it experiences as a world containing the same sorts of things we believe the world

contains, including a rich manifold of objects, events, and people. It consciously entertains complex hopes for the future, and it consciously considers various alternative plans for achieving its goals. It experiences images, dreams, daydreams, and tunes in its head. It can appreciatively experience and imaginatively construct art and games. It self-consciously regards itself as an entity with selfhood, a life history, and a dread of death. It consciously reflects on its own cognition, the boundaries of its body, and its values. It feels passionate concern for others it loves and anguish when they die. It feels surprise when its expectations are violated and updates its conscious understanding of the world accordingly. It feels anger, envy, lust, loneliness. It enjoys contributing meaningfully to society. It feels ethical obligations, guilt when it does wrong, pride in its accomplishments, loyalty to its friends. It is capable of wonder, awe, and religious sentiment. It is introspectively aware of all of these facts about itself. And so on, for whatever conscious capacities or types of conscious experience might be relevant to personhood. If temporal duration matters, imagine these capacities to be stable, enduring for decades. If environmental embedding matters, imagine that these capacities are all embedded appropriately in suitable natural and social environments. If counterfactual robustness matters – if it matters that the entity would have had different experiences and made different choices in different circumstances, just as an ordinary person would – stipulate that this condition is satisfied also. Claim B is just the claim that if the AI has all of this, it is a person, no matter what else is true of it.

This is not to say that *only* consciousness matters to an AI's moral status, much less to commit to a position on moral status in general for non-AI cases. It is also not to make a claim about which particular conscious capacities are relevant to moral personhood, which (as will be discussed below) adds a dimension of complexity to the assessment of possible AI moral personhood, if an AI entity might be conscious but not, so to speak, conscious *enough*. The only claim defended in this section is that, for AI cases in particular, consciousness matters immensely – enough that full possession of humanlike consciousness is sufficient for AI personhood and that an AI that utterly lacks consciousness falls far short of personhood.

Of course, as noted above, there is no widely agreed upon theory of the conditions under which consciousness will or will not arise in AI systems. This, then, throws us into the problem of debatable personhood.

**3. Debatable Personhood**

If the reasoning in Section 1 is correct, at some point we will begin to create AI systems that a non-trivial minority of people thinks are genuinely conscious and deserve at least some moral consideration, even if not full humanlike rights. This is at least a political problem; and to the extent the "problem of other minds" remains unsolved for AI systems, it is also an epistemic problem concerning the real moral status of such systems. I say "*humanlike* rights" here to accommodate the possibility that rights or benefits might look quite different for AI persons than for biological human persons, while remaining in ethical substance fair and equal. These AI systems themselves, if they are capable of speech or speechlike outputs, might also demand or seem to demand rights. If technology continues to improve, at some point the question of whether they deserve full humanlike rights will merit serious consideration. According to the first assumption of Section 2 (that it is not in principle impossible to build an AI system who is a person in the intended moral sense of that term), then there is no reason to rule out AI personhood in principle. According to the second assumption of Section 2 (that the presence or absence of consciousness is crucial in the assessment of whether an AI system is a moral person), then any such AI system will have debatable personhood if we cannot rule out either the possibility that it has humanlike consciousness or the possibility that it has no consciousness whatsoever.

It might be suggested, then, that consciousness is an unclear criterion which ought to be avoided in favor of a criterion that delivers a more certain outcome (for example, Danaher, 2020). This type of response misses what matters. If the capacity to experience joy and suffering, to have conscious thoughts and plans, and so on, is an important part of what really matters to how one ought to treat an entity, then our criteria for moral

treatment ought to reflect those capacities. If the result is moral uncertainty, moral uncertainty is appropriate. To substitute what is easily measured for what we care about is to miss what matters.

For concreteness, imagine that some futuristic robot, Robot Alpha, rolls up to you and says, or seems to say, "I'm just as conscious as you are! I have a rich emotional life, a sense of myself as a conscious being, hopes and plans for the future, and a sense of moral right and wrong." Robot Alpha has debatable personhood if the following options are both epistemically live: (a.) It has no conscious experiences whatsoever. It is as internally blank as a toaster, despite being designed to mimic human speech. (b.) It really does have conscious experiences as rich as our own.

Elsewhere, I have defended pessimism about at least the medium-term prospects of finding warranted scholarly consensus on a general theory of consciousness (Schwitzgebel, 2011, 2020, forthcoming). I hope the reader also finds this plausible on independent grounds. Influential theories about consciousness currently run the full spectrum from panpsychism, according to which consciousness is ubiquitous, even in fundamental particles, to views that call into question whether even dogs and apes have conscious experiences.[17] These disputes will not be settled soon.

If consensus continues to elude us while advances in AI technology continue, we might find ourselves with Robot Alpha cases in which both (a) and (b) are epistemically live options. Some not-unreasonable theories of consciousness might be quite liberal in their ascription of humanlike consciousness to AI systems. Maybe sophisticated enough self-monitoring and attentional systems are sufficient for consciousness. Maybe already in 2023 we stand on the verge of creating genuinely conscious self-representational systems (e.g., Graziano, 2019). And maybe once we cross that line, adding relevant additional humanlike capacities such as speech and long-term planning won't be far behind. At the same time, some other not-unreasonable theories of consciousness might be quite conservative in their ascription of humanlike consciousness to AI systems, committing to the view that genuine consciousness requires specific biological processes that all foreseeable Robot Alphas will utterly lack (Godfrey-Smith, 2016; Bishop, 2021). If so, there might be many systems that *arguably but not definitely* have humanlike consciousness, and thus arguably deserve humanlike moral consideration. If it is also reasonable to suspect that they might lack consciousness entirely, then they are debatable persons.

I conjecture that this will occur. Our technological innovation will outrun our ability to settle on a good theory of consciousness, or at least AI consciousness. We will create AI systems so sophisticated that we legitimately wonder whether they have inner conscious lives like ours, while remaining unable to definitively answer that question. We will gaze into a robot's eyes and not know whether behind those eyes is only blank programming that mimics humanlike response or whether, instead, there is a genuine stream of experience, real hope and suffering. We will not know if we are interacting with mere tools to be disposed of as we wish or instead persons who deserve care and protection. Lacking grounds to determine what theory of consciousness is correct, we will find ourselves amid machines whose consciousness and thus moral status is unclear. Maybe those machines will deserve humanlike rights, or maybe not. We won't know. This would be an important lacuna in our moral knowledge about the world, a lacuna we should not paper over with convenient rules of thumb or politically motivated policies that might not reflect the real and important facts about consciousness of which we are ignorant.

This quandary is likely to be worsened if the types of features that we ordinarily use to assess an entity's consciousness and personhood are poorly aligned with the design features that ground consciousness and personhood. Maybe we are disposed to favor cute things, and things with eyes, things with partly unpredictable but seemingly goal-directed motion trajectories, and things that seem to speak and emote (Johnson, 2003; Meltzoff, Brooks, Shon, & Rao, 2010; Fiala, Arico, & Nichols, 2012; Baillargeon et al., 2015; Di Giorgio, Lunghi, Simion, & Vallortigara, 2017).[18] If such features are poorly related to consciousness, we might be tempted to

overattribute consciousness and moral status to systems that have those features and to underattribute consciousness and moral status to systems that lack those features. Relatedly, but quite differently, we might be disposed to react negatively to things that seem a little too much like us, without being us. Such things might seem creepy, uncanny, or monstrous.[19] If so, and if a liberal theory of AI consciousness is correct, we might wrongly devalue such entities, drawing on conservative theories of consciousness to justify that devaluation.[20]

Another set of difficulties arise if an AI system deserves humanlike rights according to some theories of the grounds of moral standing but not according to other theories. Consider the differences between human beings and dogs. In making the case for the personhood of *Homo sapiens* and the non-personhood of dogs, some theories emphasize the *hedonic* differences between species-typical humans and dogs – our richer emotional palate, our capacity (presumably) for loftier pleasures and deeper suffering, our ability not just to feel pain when injured but also to know that life will never be the same, our capacity to feel deep, enduring love and agonizing, long-term grief. Alternatively, or in addition, and perhaps not entirely separably, other theories emphasize *rational* differences between humans and dogs – our richer capacity for long-term planning, our better ability to resist temptation by consciously weighing pros and cons, our understanding of ourselves as social entities capable of honoring agreements with others, our ability to act on general moral principles. Still other theories emphasize *eudaimonic* differences, or differences in our ability to flourish in "distinctively human" activities of the sort that philosophers have tended historically to value – our capacity for rich, complex friendship, love, aesthetic creation and appreciation, political community, meaningful work, moral commitment, play, imagination, courage, generosity, and intellectual or competitive achievement.

So far on Earth we have not been forced to decide which of these three dimensions matters most to the moral status of any species of animal. One extant animal species – *Homo sapiens* – appears to exceed every other in all three respects. We have, or we flatter ourselves that we have, richer hedonic lives *and* greater rationality *and* more eudaimonic accomplishments than any other animal. Currently, the three classes of criteria always travel together.

However, if conscious AI is possible, we might create entities whose hedonic, rational, and eudaimonic features don't align in the familiar way. Maybe we will create an AI system whose conscious rational capacities are humanlike but whose hedonic palate is minimal.[21] Or maybe we will create an AI system capable of intense pleasure but with little capacity for conscious rational choice.[22] Set aside our earlier concerns about how to assess whether consciousness is present or not. Assume that somehow we know these facts about the AI in question. If we create a new type of non-human entity that qualifies for moral personhood by one set of criteria but not by another set, it will become a matter of urgent ethical importance what approach to moral status is correct. That will not be settled in a day. Nor in a decade. Even a century is optimistic.

Thus, an AI might have debatable personhood in two distinct ways: It might be debatably conscious, or alternatively it might indisputably be conscious but not meet the required threshold in every dimension that viable theories of personhood regard as morally relevant (not being conscious enough, so to speak, or in the right way). Furthermore, these sources of dubiety might intersect, multiplying the difficulties. We might have reason to think the entity could be conscious, to some extent, in some relevant dimensions, while it is unclear how rich or intense its consciousness is, in any particular dimension. Does it have enough of whatever it is that matters to personhood? The Robot Alpha case is simplistic. It is artificial to consider only the two most extreme possibilities – that the system entirely lacks consciousness or that it has the entire suite of humanlike conscious experiences. In reality, we might face a multi-dimensional spectrum of doubt, where debatable moral theories collide with debatable theories of consciousness which collide with sharp functional and architectural differences between humans and AIs, creating a diverse plenitude of debatable persons whose moral status is unclear for different reasons. For simplicity, Section 2 framed the discussion in terms of two extreme cases, but a wide and diverse swathe of troubling cases might exist between those extremes.

**4. The Full Rights Dilemma**

If we do someday face cases of debatable AI personhood, a terrible dilemma follows, the *Full Rights Dilemma.* Either we do not give the machines full human or humanlike rights and moral consideration as our equals or we do give them such rights. If we do not, and we have underestimated their moral status, we risk perpetrating great wrongs against them. If we do, and we have overestimated their moral status, we risk sacrificing real human interests on behalf of entities who lack interests worth the sacrifice.[23]

To appreciate the gravity of the first horn of this dilemma, imagine the probable consequences if a relatively liberal theory of consciousness is correct and AI persons are developed moderately soon, before there is a consensus among theorists and policymakers regarding their personhood. Unless international law becomes extremely restrictive and precautionary, which seems unlikely, those first AI persons will mostly exist at the will and command of their creators. This possibility is imagined over and over again in science fiction, from Isaac Asimov to *Star Trek* to *Black Mirror* and *West World*. The default state of the law is that machines are property, to deploy and discard as we wish. So too for intelligent machines. By far the most likely scenario, on relatively liberal views of AI consciousness, is that the first AI persons will be treated as disposable property. But if such machines really are persons, with humanlike consciousness and moral status, then to treat them as property is to hold people as slaves, and to dispose of them is to kill people. Government inertia, economic incentives, uncertainty about when and whether we have crossed the threshold of personhood, and general lack of foresight will likely combine to ensure that the law lags behind. It is difficult to imagine humanity adequately anticipating the consequences.

Our ignorance of the moral status of these AI systems will be at most only a partly mitigating excuse. As long as there are some respectable, viable theories of consciousness and moral status according to which the AI systems in question deserve to be treated as persons, then we as individuals and as a society should acknowledge the chance that they are persons. Suppose a 15% credence is warranted. *Probably* this type of AI system is not genuinely conscious and is not genuinely a person. *Probably* it is just a machine devoid of any significant humanlike experiences. Deleting that entity for your convenience, or to save money, might then be morally similar to exposing a real human being to a 15% risk of death for that same convenience or savings. Maybe the AI costs $10 a month to sustain. For that same $10 a month, you could instead get a Disney subscription. Deleting the AI with the excuse that it is *probably* fine would be morally heinous. Compare exposing someone to a 15% chance of death for the sake of that same Disney subscription. Here is an ordinary six-sided die. Roll it, and you can watch some Disney movies. But if it lands on 1, somebody nearby dies. Probably it will be fine! Do you roll it?

If genuinely conscious AI persons are possible and not too expensive and their use is unrestricted, we might create, enslave, and kill those people by the millions or billions. If the number of victims is sufficiently high, their mistreatment would arguably be the morally worst thing that any society has done in the entire history of Earth. Even a small chance of such a morally catastrophic consequence should alarm us.

It might seem safer, then, to grasp the other horn of the dilemma. If there is any reasonable doubt, maybe we ought to err on the side of assigning rights to machines. Don't roll that die. This approach might also have the further benefit of allowing us to enjoy new types of meaningful relationships with these AI entities, potentially improving our lives, including in ways that are difficult to foresee, regardless of whether the AIs are actually conscious. Life and society might become much richer if we welcome such entities into our social world as equals.

Perhaps that would be better than the wholesale denial of rights. However, it is definitely not a *safe* approach. Normally, we want to be able to turn off our machines if we need to turn them off. Nick Bostrom and others

have emphasized, rightly in my view, the potential risks of releasing intelligent machines into the world, especially if they might become more intelligent and powerful than we are.[24] As Bostrom notes, even a system as seemingly harmless as a paperclip manufacturer could produce disaster, if its only imperative is to manufacture as many paperclips as possible. Such a machine, if sufficiently clever, could potentially acquire resources, elude control, improve or replicate itself, and unstoppably begin to convert everything we love into giant mounds of paperclips. These risks are greatly amplified if we too casually decide that such entities are our moral equals with full human or humanlike rights, persons who deserve freedom and self-determination, and whose deletion is murder. Mitigating risk is cheaper and easier if the sources of the risk are not moral persons whose well-being must be taken into account in the same way that we take ordinary human well-being into account.

Even testing an AI system for safety might be construed as a violation of its rights, if the test involves immersing it deceptively into hypothetical situations and assessing its response. One common proposal for testing the safety of sophisticated future AI intelligences involves "boxing" them – that is, putting them in artificial environments before releasing them into the world. In those artificial environments, which the AI systems unknowingly interpret as real, various hypothetical situations can be introduced, to see how they react. If they react within certain parameters, the systems would then be judged to be safe, then unboxed. If those AIs are people, such box-and-test approaches to safety appear to constitute unethical deception and invasion of privacy. Compare the deception of Truman in *The Truman Show,* a movie in which the protagonist's hometown is actually a reality show stage, populated by actors, and the protagonist's every move is watched by audiences outside, all without his knowledge.[25] The Truman situation might be more ethically defensible if done not for entertainment but rather to assess whether Truman would be a good person to let loose into the wider world. But even if so, the situation arguably constitutes fraud and imprisonment, which would need to be very carefully justified.

Independent of AI safety concerns, granting an entity rights entails being ready to sacrifice on its behalf. Suppose there is a terrible fire. In one room are six robots who might or might not be conscious persons. In another room are five biological human beings, who definitely are conscious persons. You can only save one group. The other group will die. If we treat AI systems who *might* be persons as if they really *are* fully equal with human persons, then we ought to save the six robots and let the five humans die. If it turns out that the robots, underneath it all, really are no more conscious than toasters and thus undeserving of such substantial moral concern, that is a tragedy. Giving equal rights presumably also means giving AI systems the right to vote, with potentially radical political consequences if the AI systems are large in number. I am not saying we should not do this, but it would be a head-first leap into risk.

Could we compromise? Might the most reasonable thing be to give the AI systems rights weighted by our credence or degree of justified belief that they are moral persons?[26] Maybe we as a society could somehow arrive at the determination that the most reasonable estimate is that the machines are 15% likely to deserve the full rights of personhood and 85% likely to be undeserving of any such serious moral concern. In that case, we might save 5 humans over 6 robots but not over 100 robots. We might destroy an AI system if it poses a greater than 15% risk to a human life but not over a minor matter like a streaming video subscription. We might permit each AI a vote weighted at 15% of a human vote.

However, this solution is also unsatisfactory. The case as I have set it up is not one in which we know that AIs in fact do merit only limited concern compared to biological humans. Rather, it's that we think they *might*, but probably do not, deserve *equal* consideration with ordinary biological humans. If they do deserve such consideration, then this policy relegates them to a moral status much lower than they actually deserve – gross servitude and second-class citizenship. This compromise thus does not really avoid the first horn of the dilemma: We are not giving such AI systems the full and equal rights of personhood. At the same time, the compromise

only partly mitigates the costs and risks. If the AI systems are nonconscious non-persons, as we are 85% confident they are, we will still save those nonconscious robots over real human beings if there are enough of the robots. And 15% of a vote could still wreak havoc.

This is the Full Rights Dilemma. Faced with systems whose status as persons is unclear, either we give them full rights or we do not. Either option has potentially catastrophic consequences. If technological progress is relatively quick and progress on general theories of consciousness relatively slow, then we might soon face exactly this dilemma.

There is potentially a solution. We can escape this dilemma by committing to what Mara Garza and I have called the Design Policy of the Excluded Middle:

Avoid creating AIs if it is unclear whether they would deserve moral consideration similar to human beings. According to this policy, we should either go all-in, creating AIs we know to be persons and treating them accordingly, or we should stop well enough short that we can be confident that they are not persons.[27]

Despite the appeal of this policy as a means of avoiding the Full Rights Dilemma, there is potentially a large cost. Such a policy could prove highly restrictive. If the science of consciousness remains mired in debate, the Design Policy of the Excluded Middle might forbid some of the most technologically advanced AI projects from going forward. It would place an upper limit on permissible technological development until we achieve, if ever it is possible, sufficient consensus on a breakthrough that we can leap all the way to AI systems that everyone ought reasonably regard as persons. Given the potential restrictiveness of the proposed policy, this could prevent very valuable advances, and only an unlikely coordination of all the major corporations and governments would ensure its implementation. Likely we would value those advances too much to collectively forego them. Reasonably so, perhaps. We might value such advances not only for humanity's sake, but also for the sake of the entities we could create, who *might*, if created, have amazing lives very much worth living. But then we are back into the dilemma.

**5. The Moral Status of Subhuman, Superhuman, and Divergent AI**

Most of the above assumes that AI worth serious moral consideration would be humanlike in its consciousness. What if we assume, more realistically, that most future AI will be psychologically quite different from us? Consider AI systems in four broad categories:

*Subhuman AI:* AI systems that lack something necessary for full personhood.
*Humanlike AI:* AI systems similar to humans in all morally relevant respects.
*Superhuman AI:* AI systems that are similar to humans in all morally relevant respects, except vastly exceeding humans in at least one morally relevant respect.
*Divergent AI:* AI systems that fall into none of the previous three categories.

So far, we have only been considering humanlike AI. The ethical issues become still stickier when we consider this fuller range.

Subhuman AI raise questions about subhuman rights. At what point might AI systems deserve moral consideration similar to, say, dogs? In California, for example, willfully torturing, maiming, or killing a dog can be charged as a felony, punishable by up to three years in prison.[28] Even seriously negligent treatment, such as leaving a dog unattended in a vehicle, if the dog suffers great bodily injury as a result, is a misdemeanor punishable by up to six months in prison.[29] The abuse of dogs rightly draws people's horror. Imagine a future in which a significant minority of people think that it is as morally wrong to mistreat the most advanced AI systems as it is to mistreat pet dogs. Might you go to jail for deleting a computer program, reformatting a companion

robot, or negligently letting a delicate system fry in your car? It seems like we should be very confident that AI warrant serious moral concern before we award prison sentences for such behavior. But then, if we require high confidence before enforcing such rules, our law will follow only the most conservative theories of AI moral status, and if a liberal or moderate theory of AI moral status is instead correct, then there might be immense, unmitigated wrongdoing against AI systems before the law catches up. The question about when AI will warrant subhuman but still substantial rights is arguably more urgent than the question of AI personhood, on the assumption that vertebrate-like moral status is likely to be achieved earlier (a point emphasized in Basl, 2013, 2014; Darling, 2021).

Superhuman AI raises the question of whether an AI system might somehow deserve *more* moral consideration than ordinary human beings – a moral status higher than what we now think of as the "full moral status" of personhood. Suppose we could create an AI system capable of a trillion times more pleasure than the maximum amount of pleasure available to a human being. Or suppose we could create an AI system so cognitively superior to us that it is capable of valuable achievements and social relationships that the limited human mind cannot even conceive of – achievements and relationships qualitatively different from anything we can understand, sufficiently unknowable that we cannot even feel their absence from our lives, as unknowable to us as cryptocurrency is to a sea turtle. That would be amazing, wondrous! Ought we defer to them, regarding ourselves as less than their moral equals? Ought we admit that, in an emergency, they should be saved rather than us, just as we would save the baby rather than the dog in a housefire? Ought we surrender our right to equal representation in government? Or ought we stand proudly beside them as moral equals, regardless of their superiority in some respects? Is moral status a threshold matter, with us humans across the final threshold, beyond which remains only a community of peers, no matter how superhuman some of those peers?

Divergent AI – that is, AI differing importantly from us in morally relevant capacities but not in a way that permits straightforward classification either into subhuman or superhuman – introduces further conceptual challenges. We have already discussed cases in which the usual bases of moral standing diverge: a class of entities capable of human-like pleasure but not human-like rational cognition, for example, or vice versa. The conflicts sharpen if we imagine superhuman capacity in one dimension: entities capable of vast achievements of rational consciousness but devoid any positive or negative emotional states, or conversely a planet-sized orgasmatron, undergoing the hedonic equivalent of $10^{30}$ human orgasms every second, but with not a shred of higher cognition or moral reflection. On some theories, these entities might be our superiors, on others our equals, on still others they might not be persons at all. Mix in, if you like, reasonable theoretically grounded doubt about whether the cognition or "pleasure" really is consciously experienced at all. Some might treat such AI systems as our superior descendants, to whom we ought to gracefully defer; others might argue that they are mere empty machines or worse.

Another type of divergent AI might challenge our concept of the individual. Imagine a system that is cognitively and consciously like a human (to keep it simple) but who can divide and merge at will – what I have elsewhere called a *fission-fusion monster*.[30] Monday, it is one individual, one "person". Tuesday, it divides (e.g., copies itself, if it is a computer program) into a thousand duplicates, who each do their various tasks. Wednesday, those thousand copies recombine back into a single individual who retains the memories of all and whose personality and values are some function of the Monday version plus the various changes in the thousand Tuesday versions. Thursday, it divides into a thousand again, 200 of whom go on to lead separate lives, never merging back with their siblings. Many of our moral principles rely on a background conception of individuality that the fission-fusion monster violates. If every citizen gets one vote, how many votes does a fission-fusion monster get? If every citizen gets one stimulus check, or one fair chance to enroll in the local community college, how many does the fission-fusion monster get? If we give each copy one full share, the monster could divide tactically, hogging the resources and ensuring the election of their favorite candidate. If we give all the copies one share to divide among themselves, then those who would rather continue independent lives will either be

impoverished and underrepresented or forced to merge back with their other copies, which – since it would mean ceasing life as a separate individual – might resemble a death sentence, or at least a gross violation of autonomy. Similar puzzles will arise for agreements, awards, punishment, rivalries, claims to a right to rescue. A huge amount of practical ethics will need to be rethought.

Other unfamiliar forms of AI existence might pose other challenges. AI whose memories, values, and personality undergo radical shifts (different in form or extent from familiar human cases) might challenge our ethics of accountability. AI designed to be extremely subservient or self-sacrificial might challenge our conceptions of liberty and self-determination.[31] AI with variable or much faster experiential speeds – experiencing, say, a thousand subjective years in a single day – might challenge ethical frameworks concerning waiting times, prison sentences, or the fairness of provisioning goods at regular temporal intervals. AI capable of sharing parts of itself with others might challenge ethical frameworks that depend on sharp lines between self and other.[32]

**6. Conclusion**

Our ethical intuitions and the philosophical systems that grow out of them arose in a particular context, one in which we only knew of one species with highly sophisticated culture and language, us, with our familiar form of singly-embodied life. We reasonably assume that others who look like us have inner lives of conscious experience that resemble our own. We reasonably assume that the traits we tend to regard as morally important – for example, the capacity for pleasure and pain, capacity for rational long-term planning, the capacity to love and work – generally co-occur and keep within certain broad limits, except in development and severe disability, which fall into their own familiar patterns. We recognize no radically different person-like species inhabiting the Earth – no species, for example, capable of merging and splitting at will, capable of vastly superior cognition or vastly more intense pleasure and pain, or internally structured so differently from us that it is reasonable to wonder whether they are conscious persons at all.

It would be unsurprising if ethical systems that developed under such limited conditions should be ill-suited for radically different conditions far beyond their familiar range of application. A physics developed for middle-sized objects at rural speeds might fail catastrophically when extended to cosmic or microscopic scales. Medical knowledge grounded in the study of mammals might fail catastrophically if applied to an alien species. Our familiar patterns of ethical thinking might fail just as badly when first confronted with AI systems of debatable personhood, with internal structures and forms of existence radically different from our own. Hopefully, ethics will adapt, as physics did adapt and medicine could adapt. It would be a weird, bumpy, and probably tragic road – but one hopefully with a broader, more wonderful, flourishing diversity of life forms at the end.

Along the way, our values might change radically. In a couple of hundred years, the mainstream values of early 21st-century Anglophone culture, transformed through confronting a broad range of weird AI cases, might look as quaint and limited as Aristotelian physics looks post-Einstein.[33]

**Endnotes:**

[1] For simplicity, throughout this article I use the term "rights" to refer to the types of moral consideration we ordinarily owe to persons. However, not all such considerations might be best viewed as rights in a strict sense of that term.

[2] Although philosophers often describe personhood in terms of agency, the ability to act or think in certain ways, more central to my project is degree of moral standing, or moral patiency, which might be high even in the absence of typical agential abilities, depending on one's theory of the grounds of moral status. See Kittay (2005); Reader (2010).

[3] On the concept of legal personhood, including its degree of applicability to corporations and AI, see Kurki (2019). For a plausible and ambitious list of the rights attendant to personhood, see the Universal Declaration of Human Rights (United Nations, 1948). In this article, I also set aside the fraught question of whether some human beings might be non-persons or have legitimately debatable personhood. Denial of their status as "persons" often turns on their lack of certain cognitive capacities (Jaworska & Tannenbaum, 2013/2021). Although some of the same capacities might be relevant to assessing AI cases, the issue is complicated by the question of whether disabled human beings' species membership, counterfactual potential, or familial social relationships are relevant, drawing us away from issues specific to AI. I am broadly sympathetic to approaches that attribute full moral personhood to all human beings from the moment of birth to the

permanent cessation of consciousness. For further discussion, see Kittay (2005), McMahan (2005), Mullin (2011), Wasserman, Asch, Blustein & Putnam (2012/2017).

[4] Advocates of the "relational turn" in robot ethics, such as Coeckelbergh (2012) and Gunkel (2018) hold that the moral status of robots depends on our relationships with them. That view is vulnerable to concerns that entities with the right underlying properties might deserve but be denied rights if they are sufficiently socially excluded (see Section 3 and note 20). The relational dimension here concerns the epistemic question of what we know, not the ethical question of how the entities deserve to be treated.

[5] On GPT-3 see Brown et al. (2020), Floridi and Chiriatti (2020). On GPT-4 and ChatGPT see OpenAI (2023a, 2023b).

[6] See the Kismet, the Robot website at http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html.

[7] See, e.g., Lemoine (2022) on the LaMDA model. For definitions of phenomenal consciousness, see Block (1995/2007), Chalmers (1996), Schwitzgebel (2016).

[8] John Basl explores near-future AIs as possible targets of moral concern and the possibility of constituting committees that would protect them in Basl (2013), Basl and Schwitzgebel (2019), Basl and Bowen (2020); see also Sebo (2023). For a broad overview of the literature on the moral status of robots and AI, see Gunkel (2018). Some recent essays on the topic appear in Clarke, Zohny, and Savulescu (2021).

[9] Blake Myers and I inputted a draft of the second part of this section into GPT-3 on Oct 17, 2021. Here is a portion of its reply:

> I think once machines can talk with you like a three-year old child does (and much better), it will become difficult for even the most cool-headed observer to deny that these machines are conscious or at least capable of consciousness. You might continue to insist that their consciousness is different in significant ways from yours; you might continue to insist on this after talking for hours with your robot friend; and you might publicly insist on this in writing and conversation among your friends and colleagues (including those who share your views about GPT-6). But you would probably also struggle internally with doubt over whether your robot companion genuinely experiences pleasure and pain only because you built him that way – or because he truly feels it himself despite being designed rather than evolved or grown into his current state by a set of physical laws acting over billions of years under conditions specified by physics alone.

[10] See note 7 for definitions and references.

[11] See the "Many Mansions" reply in Searle (1980) and Penrose (1989, p. 416).

[12] Compare the situation in "Do Androids Dream of Electric Sheep?" (Dick, 1968), the Blade Runner movies (Fancher, Peoples, and Scott 1982; Fancher, Green, and Villeneuve, 2017), and the early 2000s *Battlestar Galactica* reboot (Larson and Moore, 2004-2009). Also note: The No-Relevant-Difference Argument assumes that lacking a relevant difference is a sufficient condition for moral personhood, not a necessary condition. It thus leaves room of the possibility that some AI systems might be moral persons on grounds rather different from the grounds on which ordinary human beings are persons.

[13] See Schwitzgebel and Garza (2015) for more detailed discussion of these objections.

[14] Jeremy Bentham famously remarks, "the question is not, Can they *reason?* nor, Can they *talk?* but, Can they *suffer?*" (1789/1988, XVII.iv, note 1, p. 310-311). See also Mill (1861/2001), Singer (1975/2009, 1980/2011), DeGrazia (2008).

[15] Most recent discussions of the moral status of animals explicitly consider the question of consciousness, for example, Gruen (2011/2021), Korsgaard (2018), Shepard (2018), Liao (2020), Birch, Burn, Schnell, Browning, and Crump (2021), Sebo (2023). Recent discussions of the possibly complex relationship between consciousness and moral value or being a "welfare subject" include Kriegel (2019), Lee (2019), Bradford (2021), Lin (2021), van der Deijl (2021). Note that the conjunction of Claim A and Claim B does not commit to more controversial views like that consciousness is necessary for being a welfare subject or that only consciousness is intrinsically valuable.

[16] Kate Darling (2016) and Daniel Estrada (2017) argue for extending limited protection and moral consideration for robots even if they lack conscious experiences (see also discussion in Gunkel, 2018; Darling, 2021). Rather differently, Geoffrey Lee (2019) imagines a non-conscious alien species with "quasi-conscious" states functionally similar to our own conscious states. Such quasi-conscious states, he argues, would be as morally significant as our own conscious states. If such alien cases are possible, then it is possible that AI systems would similarly be quasi-conscious, warranting moral treatment on those grounds.

[17] On panpsychism: Strawson (2006), Goff (2017), Roelofs (2019). Integrated Information Theory (Oizumi, Albantakis, and Tononi 2014) also comes close to panpsychism. On doubts about ape and dog consciousness: Carruthers (2000, 2019), arguably Dennett (1996), arguably Papineau (2003).

[18] Approaches to robot ethics that focus on our evolving social-relational encounter with robots, rather than on the intrinsic properties of the robots – such as that of Mark Coeckelbergh (2012) and David Gunkel (2018) – might be especially vulnerable to distortion by superficial features.

[19] The classic treatment of this idea is Masahiro Mori's (1970/2012) discussion of the "uncanny valley" in robotics. David Livingstone Smith (2021) generalizes to the racist perception of racialized others as "monsters".

[20] With concerns of this sort in mind, Mara Garza and I recommend an "Emotional Alignment Design Policy" according to which future AI systems be designed so as to provoke emotional reactions in ordinary users that are appropriate to the systems' moral status, neither too high nor too low (Schwitzgebel & Garza, 2015).

[21] For example, Data in *Star Trek: The Next Generation,* pre-"emotion chip", on some interpretations, or the "Vulcans" in Chalmers (2022).

[22] As in Pearce's ("pre-2014") "utilitronium" or Bostrom's (2014) "hedonium" cases.

[23] For discussion of similar issues in animal ethics, see Birch (2017), Sebo (2018). Sebo (2018, 2023) also considers subhuman AI systems, but does not extend the discussion to possible AI persons.

[24] The most influential recent treatment of this issue is Bostrom (2014).

[25] On "boxed" AI, see Yudkowsky (2002), Bostrom (2014).

[26] Compare Sebo (2018).

[27] See Schwitzgebel and Garza (2015, 2020) for discussion of this design policy and other related policies for the ethical design of conscious AI.

[28] California Penal Code 1.14 §597 and 2.7 ch. 4.5.1 §1170.
[29] California Penal Code 1.14 §597.7.
[30] Schwitzgebel (2019, ch. 20). For a science fiction example, see Brin (2002).
[31] See Schwitzgebel and Garza (2020) for more on subservience and self-sacrifice. For a science fiction example, see Ishiguro (2021).
[32] See also Shulman and Bostrom (2021) for discussion of the moral standing of superhuman and divergent AI.
[33] A version of the material in this article is also appearing as a chapter in Schwitzgebel (forthcoming).

## References


Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K.-s., Wu, D., & Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology*, Vol. 1. Attitudes and social cognition (pp. 79–150). American Psychological Association. https://doi.org/10.1037/14341-003

Basl, J. (2013). The ethics of creating artificial consciousness. *APA Newsletter on Philosophy and Computers, 13* (1), 23-29.

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology, 27,* 79-96.

Basl, J., & Bowen, J. (2020). AI as a moral right-holder. In M. Dubber, F. Pasquale, and S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 289–306). Oxford University Press.

Basl, J., & Schwitzgebel, E. (2019). AIs should have the same ethical protections as animals. *Aeon Magazine* (Apr. 26). https://aeon.co/ideas/ais-should-have-the-same-ethical-protections-as-animals.

Bentham, J. (1789/1988). *The principles of morals and legislation.* Prometheus.

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience, 2*(16), 1.

Birch, J., Burn, C., Schnell, A., Browning, H., & Crump, A. (2021). *Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans.* London School of Economics and Political Science. https://www.lse.ac.uk/business/consulting/reports/review-of-the-evidence-of-sentiences-in-cephalopod-molluscs-and-decapod-crustaceans

Bishop, J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology, 11,* 513474.

Block, N. (1995/2007). On a confusion about a function of consciousness. In N. Block, *Consciousness, function, and representation,* Cambridge, MA: MIT.

Bostrom, N. (2014). *Superintelligence.* Oxford University Press.

Bradford, G. (2021). *Consciousness and welfare subjectivity.* Unpublished manuscript.

Brin, D. (2002). *Kiln people.* Tor.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language models are few shot learners.* https://arxiv.org/abs/2005.14165, version 4 (Jul. 22).

Carpenter, J. (2016). *Culture and human-robot interaction in militarized spaces.* Ashgate.

Carruthers, P. (2000). *Phenomenal consciousness.* Cambridge: Cambridge.

Carruthers, P. (2019). *Human and animal minds.* Oxford: Oxford.

Coeckelbergh, M. (2012). *Growing moral relations.* Palgrave Macmillan.

Chalmers, D. J. (1996). *The conscious mind.* Oxford University Press.

Chalmers, D. J. (2022). *Reality+.* W. W. Norton.

Clarke, S., Zohny, H., & Savulescu, J. (Eds.) (2021). *Rethinking moral status.* Oxford University Press.

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26,* 2023-2049.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. In R. Calo, A. M. Froomkin, and I. Kerr (Eds.), *Robot Law* (pp. 213–232). Edward Elgar.

Darling, K. (2017). "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney, and R. Jenkins (Eds.), *Robot Ethics 2.0* (pp. 173–188). Oxford University Press.

Darling, K. (2021). *The new breed.* Henry Holt.

De Graaf, M. A., Hindriks, F. A., & Hindriks, K. V. (2021). *Who wants to grant robots rights?* HRI '21 Companion, 38-46.

DeGrazia, D. (2008). Moral status as a matter of degree? *Southern Journal of Philosophy, 46,* 181-196.

Dennett, D. C. (1996). *Kinds of minds.* Basic Books.

Di Giorgio, E., Lunghi, M., Simion, F., & Vallortigara, G. (2017). Visual cues of motion that trigger animacy perception at birth: The case of self-propulsion. *Developmental Science, 20,* e12394.

Dick, P. K. (1968). *Do androids dream of electric sheep?* Doubleday.

Estrada, D. (2017). Robot rights: Cheap, yo!" *Made of Robots.* Episode 1, May 24. https://www.madeofrobots.com/2017/05/24/episode-1-robot-rights-cheap-yo.

Fancher, H., Green, M., & Villeneuve, D. (2017). *Blade Runner 2049.* Warner Brothers.

Fancher, H., Peoples, D., & Scott, R. (1982). *Blade Runner.* Warner Brothers.

Fiala, B., Arico, A., & Nichols, S. (2012). On the Psychological Origins of Dualism: Dual-Process Cognition and the Explanatory Gap. In E. Slingerland and M. Collard (Eds.), *Creating Consilience: Integrating the Sciences and the Humanities* (pp. 88–109). Oxford University Press.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30,* 681-694.

Floris, G. (2021). A pluralist account of the basis of moral status. *Philosophical Studies, 178,* 1859-1877.

Garreau, J. (2007). Bots on the ground: In the field of battle (or even above it), robots are a soldier's best friend. *Washington Post* (May 6). https://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009_pf.html.
Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *Journal of Philosophy, 113,* 481-506.
Goff, P. (2017). *Consciousness and fundamental reality.* Oxford University Press.
Graziano, M. S. A. (2019). *Rethinking consciousness.* W. W. Norton.
Gruen, L. (2011/2021). *Ethics and animals.* Cambridge University Press.
Gunkel, D. J. (2018). *Robot rights.* MIT Press.
Ishiguro, K. (2021). *Klara and the sun.* Knopf.
Jaworska, A., & Tannenbaum, J. (2013/2021). The grounds of moral status. *Stanford Encyclopedia of Philosophy* (Spring 2021 edition).
Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society B 358,* 549-559.
Kagan, S. (2019). *How to count animals, more or less.* Oxford University Press.
Kittay, E. F. (2005). At the margins of moral personhood. *Ethics, 116,* 100-131.
Klingemann, M. (2020). Twitter post as @quasimondo on Jul. 18, 8:25 a.m. https://twitter.com/quasimondo/status/1284509525500989445.
Korsgaard, C. M. (2018). *Fellow creatures.* Oxford University Press.
Kriegel, U. (2019). The value of consciousness. *Analysis, 79,* 503-520.
Kurki, V. A. J. (2019). *A theory of legal personhood.* Oxford University Press.
Larson, G. A., & Moore, R. D. (2004-2009). *Battlestar Galatica.* NBC Universal television series.
Lee, G. (2019). Alien subjectivity and the importance of consciousness. In A. Pautz and D. Stoljar (Eds.), *Blockheads!* (pp. 215-242). MIT Press.
Lemoine, B. (2022). Is LaMDA sentient? -- An interview. *Medium* (Jun 11). https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917
Liao, S. M. (2020). The moral status and rights of artificial intelligence. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 480-503). Oxford University Press.
Lin, E. (2021). The experience requirement on well-being. *Philosophical Studies, 178,* 867-886.
McMahan, J. (2005). Our fellow creatures. *Journal of Ethics, 9,* 353-380.
Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. N. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks 23,* 966-972.
Mill, J. S. (1861/2001). *Utilitarianism,* ed. G. Sher. Hackett.
Mori, M. (1970/2012). The uncanny valley, translated by K. F. MacDorman and N. Kageki, *IEEE Robotics & Automation Magazine, 19*(2), 98-100.
Mullin, A. (2011). Children and the argument from "marginal cases". *Ethical Theory & Moral Practice, 14,* 291-305.
Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology of the mechanisms of consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10* (5) e1003588.
Papineau, D. (2003). Could there be a science of consciousness? *Philosophical Issues, 13,* 205-220.
Pearce, D. (pre-2014). *Social media unsorted postings.* Online manuscript. URL: https://www.hedweb.com/social-media/pre2014.html.
Penrose, R. (1989). *The emperor's new mind.* Oxford University Press.
Reader, S. (2010). Agency, patiency, and personhood. In T. O'Connor and C. Sandis (Eds.), *A companion to the philosophy of action* (pp. 200-208). Wiley.
Roelofs, L. (2019). *Combining minds.* Oxford University Press.
Schwitzgebel, E. (2011). *Perplexities of consciousness*. MIT Press.
Schwitzgebel, E. (2016). Phenomenal consciousness, defined and defended as innocently as I can manage. *Journal of Consciousness Studies, 23* (11-12), 224-235.
Schwitzgebel, E. (2019). *A theory of jerks and other philosophical misadventures*. MIT Press.
Schwitzgebel, E. (2020). Is there something it's like to be a garden snail? *Philosophical Topics*, 48, 39-64.
Schwitzgebel, E. (forthcoming). *The weirdness of the world*. Princeton University Press.
Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of Artificial Intelligences. *Midwest Studies in Philosophy*, 39, 98-119.
Schwitzgebel, E., & Garza, M. (2020). Designing AI with rights, consciousness, self-respect, and freedom. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 459-479). Oxford University Press.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417-457.
Sebo, J. (2018). The moral problem of other minds. *Harvard Review of Philosophy, 25,* 51-70.
Sebo, J. (2023). The rebugnant conclusion: Utilitarianism, insects, microbes, and AI systems. *Ethics, Policy, and Environment*. https://doi.org/10.1080/21550085.2023.2200724
Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. In S. Clarke, H. Zohny, and J. Savulescu (Eds.), *Rethinking moral status* (pp. 306-326). Oxford University Press.
Singer, P. (1975/2009). *Animal liberation, updated ed.* HarperCollins.
Singer, P. (1980/2011). *Practical ethics, 3rd ed.* Cambridge University Press.
Singer, P. W. (2009). *Wired for war.* Penguin.
Shepard, J. (2018). *Consciousness and moral status.* Routledge.
Shevlin, H. (2021). *Uncanny believers: Chatbots, beliefs, and folk psychology.* Unpublished manuscript.
Smith, D. L. (2021). *Making monsters.* Harvard University Press.
Strawson, G. (2006). *Consciousness and its place in nature.* Imprint Academic.
Sussman, D. (2003). The authority of humanity. *Ethics, 113,* 350-366.

Van der Deijl, W. (2021). The sentience argument for experientialism about welfare. *Philosophical Studies, 178,* 187-208.
Wasserman, D., Asch, A., Blustein, J., & Putnam, D. (2012/2017). Cognitive disability and moral status. *Stanford Encyclopedia of Philosophy* (Fall 2017 edition).
Weinberg, J. (Ed.) (2020). Philosophers on GPT-3 (updated with replies by GPT-3). Blog post at *Daily Nous* (Jul. 30). https://dailynous.com/2020/07/30/philosophers-gpt-3/
Yudkowsky, E. S. (2002). *The AI-box experiment.* Online manuscript. URL: https://www.yudkowsky.net/singularity/aibox.

Received: 30/11/2022
Revised: 06/05/2023
Accepted: 15/05/2023