Moral Consideration for AI Systems by 2030[1]

Jeff Sebo[2] and Robert Long[3]

Abstract

This paper makes a simple case for extending moral consideration to some AI systems by 2030. It involves a normative premise and a descriptive premise. The normative premise is that humans morally ought to extend moral consideration to beings that have a non-negligible chance, given the evidence, of being sentient or otherwise morally significant. The descriptive premise is that some AI systems do in fact have a non-negligible chance, given the evidence, of being sentient or otherwise morally significant by 2030. The upshot is that humans have a moral duty to extend moral consideration to some AI systems by 2030. And if we have a moral duty to do that, then we plausibly also have a moral duty to start preparing to discharge that duty now, so that we can be ready to treat AI systems with respect and compassion when the time comes.

1. Introduction

AI capabilities are advancing rapidly. At the time of writing, Google, Meta, Microsoft, and other companies are racing to create and deploy AI systems that can produce novel essays, photos, videos, or other outputs based on simple written prompts (Zhang et al., 2023). These systems are already advanced, and further advances seem very likely. For instance, we might one day produce AI systems that produce intelligent behavior by making use of integrated and embodied capacities for perception, learning, memory, anticipation, social awareness, self-awareness, and reasoning, in much the same way that human and nonhuman animals do (as well as in very different kinds of ways). And at that point, AI capabilities might not only match but vastly exceed human and nonhuman animal capabilities on a wide range of tasks.

These developments raise urgent ethical questions. Some concern how AI systems might harm humans and other animals. For example, AI systems might make many jobs obsolete (Acemoglu et al., 2022; Chelliah, 2017). They might amplify racism, sexism, speciesism, and other oppressions contained within their training data (Zajko, 2021; see also: Long, 2021),

[1] Thanks to Toni Adleberg for extensive research assistance for this paper along with helpful comments and suggestions on the penultimate draft. Thanks also to Joel Becker, Christian Tarsney, Elliott Thornley, Hayden Wilkinson, and Thomas Woodward for helpful feedback or discussion.

[2] Director of the Mind, Ethics, and Policy Program, New York University; jeffsebo@gmail.com

[3] Philosophy Fellow, Center for AI Safety; rgblong@gmail.com

disproportionately impacting people with intersecting marginalized identities (Guo & Caliskan, 2021; Tan & Chelis, 2019). They might assist humans in harming each other by spreading misinformation or creating novel weapons (Longpre et al., 2022). And as their capabilities increase, these risks will increase as well, leading to scenarios where AI systems either drive humans and other animals to extinction or permanently reduce our capacity for flourishing (Vold & Harris, 2021; Bostrom, 2014; Hendryks, 2023; Singer & Tse, 2022).

Another, more neglected set of questions concerns how humans might harm AI systems. Most experts now agree that sentient beings — that is, beings who can consciously experience positive and negative states like pleasure and pain — have moral standing — that is, they merit moral consideration for their own sakes. To be clear, many experts still disagree about whether sentience is *necessary* for moral standing; some hold that, say, consciousness without sentience or agency without consciousness is sufficient (Chalmers, 2022; Delon, n.d.; Ladak, 2023). But they still agree that sentience is *sufficient* for moral standing. We thus need to ask whether and when AI systems might have morally significant features such as sentience, consciousness, and agency, and what might follow for our moral responsibilities to them.

This paper makes a simple case for extending moral consideration to some AI systems by 2030. It involves a normative premise and a descriptive premise. The normative premise is that humans morally ought to extend moral consideration to beings that have a non-negligible chance, given the evidence, of being sentient or otherwise morally significant. The descriptive premise is that some AI systems do in fact have a non-negligible chance, given the evidence, of being sentient or otherwise morally significant by 2030. The upshot is that humans have a moral duty to extend moral consideration to some AI systems by 2030. And if we have a moral duty to do that, then we plausibly also have a moral duty to start preparing to discharge that duty now, so that we can be ready to treat AI systems with respect and compassion when the time comes.

Before we begin, we should note several features of our argument and conclusion that will be relevant here. First, our discussion of both the normative premise and the descriptive premise are somewhat compressed. Our aim in this paper is not to establish either premise with maximum rigor, but rather to motivate them in clear and concise terms and then show how they interact. We think that examining these premises together is important, since while we might find each one unremarkable when we consider them in isolation, what happens when we put them together is striking: They jointly imply that we should expand our moral circle substantially, to a vast number and wide range of additional beings. We aim to show how that happens and indicate why this conclusion is more plausible than it might initially appear to be.

Second, our argument in this paper is intentionally conservative in several respects. One is that we focus on estimating when AI systems will have a non-negligible chance of being *sentient*, without considering other features that might be sufficient for moral standing. If we allowed for the possibility that, say, non-sentient agents can *also* merit moral consideration for their own sakes, then that would strengthen our case for extending moral standing to AI systems sooner rather than later. A complete examination of the question of AI moral standing would consider these possibilities as well. But our interest here is in showing that even if we accept a relatively demanding standard for moral standing, we can still expect at least some AI systems to have at least a non-negligible chance of meeting that standard within the decade.

Another respect in which our argument is intentionally conservative is this: When we develop our normative premise, we assume for the sake of argument that a non-negligible chance means a 0.1% chance or higher. And when we develop our descriptive premise, we make conservative assumptions about how demanding the requirements for sentience are and how difficult these requirements are to satisfy. Our own view is that the threshold for non-negligibility is much lower than 0.1%, and that the chance that some AI systems will be sentient by 2030 is much higher than 0.1%. But we focus on this threshold here to be as generous as possible to skeptics about our view, and to emphasize that in order to avoid our conclusion, one must take extremely bold and tendentious positions about either the values, the facts, or both.

Finally, we should emphasize that our conclusion here has no immediately obvious or straightforward implications for how humans should treat AI systems. Even if we agree that we should extend moral standing to AI systems by 2030, we need to consider many further questions before we know what that means in practice. For instance, how much do AI systems count and in what ways do they count? How do our actions and policies affect them and what do we owe them in light of these effects? And how can, and should, we make tradeoffs between humans, animals, and AI systems in practice? We will consider possible tradeoffs in more detail below. For now, we will simply note that answering these questions responsibly will take a lot of work from a lot of people, which is why we should start asking these questions now.

2. The Normative Premise

The normative basis for our argument is simple, plausible, and widely accepted: We have a moral duty to consider non-negligible risks when deciding what to do. For example, if an action or policy has a non-negligible chance of gravely harming or killing someone against their will, then that risk counts against that action or policy. Of course, the risk may or may not count *decisively* against the action or policy; that will depend on the details of the case, as well as on our further moral assumptions, some of which we can consider in a moment. But whether or not this kind of risk is a decisive factor in our decision-making, it should at least be a factor. And importantly, this can be true even if the risk is very low, for instance, even if the chance that the action or policy might harm someone against their will is only one in a thousand.

There are many examples of this phenomenon, ranging from the ordinary to the extraordinary. To take an ordinary example, many people rightly see driving drunk as wrong because it carries a non-negligible risk of leading to an accident, and because this risk clearly trumps any benefits that driving drunk may involve. Granted, we can imagine exceptions to this rule; for instance, if your child is dying, and if the only way that you can save them is by driving them to a nearby hospital while drunk, then we might or might not think that the benefits of driving drunk outweigh the risks in this case, depending on the details and our further assumptions. But in standard cases, we rightly hold that even a low risk of causing an accident is reason enough to make driving drunk wrong. And either way, the risk should at least be considered.

Alternatively, to take an extraordinary example, suppose that building a superconducting supercollider carries a non-negligible risk of creating a black hole that swallows the planet. In this case, many people would claim that this experiment is wrong because it carries this risk, and because this risk generally outweighs the benefits of scientific exploration (Greene, 2020). Again, we can imagine exceptions; for instance, if the sun will likely destroy the planet within the century, and if the only way that we can survive is by advancing particle physics, then we might think that the benefits of this experiment outweigh the risks in this case. But otherwise, we might hold that even a low risk of creating a black hole is reason enough to make the experiment wrong. And either way, the risk should once again at least be considered.

Of course, these further details often matter. For instance, suppose that one superconducting supercollider carries a one in a thousand chance of creating a black hole, whereas another superconducting supercollider carries only a one in a hundred chance of doing so. Suppose further that the black hole would be equally bad either way, causing the same amount of death and destruction for humans and other sentient beings. In this case, should we assign equal weight to these risks in our decision-making, because they both carry a non-negligible risk of creating a black hole and this outcome would be equally bad either way? Or should we instead assign more weight to the risk involved with using the second superconducting supercollider, because it carries a higher risk of creating a black hole in the first place?

According to the precautionary principle (on one interpretation), we should take the former approach. If an action or policy carries a non-negligible risk of causing harm, then we should assume that this harm will occur and ask whether the benefits of this action or policy outweigh this harm. In contrast, according to the expected value principle, we should take the latter approach. If an action or policy carries a non-negligible risk of causing harm, then we should multiply the probability of harm by the level of harm and ask whether the benefits of this action or policy outweigh the resulting amount of harm. These approaches use different methods to incorporate non-negligible risks into our decisions, but importantly for our purposes here, they do both incorporate these risks into our decisions (Sebo, 2018; Birch, 2017).

To take another example, suppose that a third superconducting supercollider carries only a negligible chance (say, a one in a quintillion chance) of creating a black hole. But suppose that, once again, the black hole would be equally bad as before, causing the same amount of death and destruction for humans and other sentient beings. Should we assign at least *some* weight to this risk in our decision-making, in spite of the fact that the probability is so low, because the risk is still present and it would still be bad if this outcome came to pass? Or should we instead assign no weight at all to this risk in our decision-making, in spite of the fact that the risk is still present and it would still be bad if this outcome came to pass, simply because the probability of harm is so low that we can simply neglect it entirely for practical purposes?

According to what we can call the no threshold view, we should take the former approach. We should consider all risks, including extremely low ones. Granted, if we combine this view with the expected value principle, then we can assign extremely little weight to extremely unlikely outcomes, all else equal. But we should still assign weight to these outcomes. In contrast, according to what we can call the low threshold view, we should take the latter approach. We should consider all non-negligible risks — that is, risks above a particular probability threshold — but we can permissibly neglect all negligible risks — that is, risks below that threshold. Of course, this view faces the question about what that threshold should be, and the implications of these views will differ more or less depending on that (Sebo, 2023; Wilkinson, 2022).

Despite these disagreements, we can all agree on this much: We should assign at least some weight to at least non-negligible risks. In what follows, we will assume that much and nothing more. As for what level of risk counts as non-negligible, philosophers generally set the threshold somewhere between one in ten thousand and one in ten quadrillion (Monton, 2019). (If a superconducting supercollider carried a one in ten thousand chance of killing all humans, we would want to know that!) But for our purposes here, we will assume that the threshold is higher than that, at *one in a thousand*. That way, when we explain how our normative assumption leads to a moral duty to extend at least some moral consideration to at least some near future AI systems, no one can reasonably accuse us of stacking the deck in favor of our conclusion.

Now, how does our assumption that we should consider non-negligible risks apply to the question of whether we should treat an AI system as sentient? The application is mostly straightforward, but with a few caveats. This is the general idea: If a being is sentient, then they can be harmed. So, if a being has a non-negligible chance of being sentient, then they have a non-negligible chance of being capable of being harmed. And, if a being has a non-negligible chance of being capable of being harmed, then moral agents have a duty to consider whether our actions might harm them for their own sake. Finally, if moral agents have a duty to consider whether our actions might harm someone for their own sake, then that means that we have a duty to treat them as having moral standing, albeit with a few important caveats.

Now, here are the caveats. First, to say that moral agents should *treat* a being as having moral standing is not to say that the being *does* have moral standing. If sentience is necessary and sufficient for moral standing and if a being has a non-negligible chance of being sentient, given the information available to us, then we should treat this being as having moral standing. But if this being is not, in fact, sentient, then this would be an example of a false positive. It would be a case where we treat a non-sentient, non-morally significant being as sentient and morally significant. False positives carry costs, and we will discuss how we should think about these costs below. But what matters for present purposes is that our argument is about whether we should *treat* AI systems as having moral standing, not whether they *do*.

A second caveat is that to say that moral agents should treat a being as having moral standing is not to say how we should treat this being all things considered. Here, a lot depends on our further assumptions. For example, if we perceive tradeoffs between what this being might need and what everyone else needs, then we of course need to consider these tradeoffs carefully. And if we accept an expected value principle and hold that a being is, say, only 10% likely to be morally significant, then we can assign their interests only 10% of the weight we otherwise would, all else equal. We will consider these points below as well. But what matters for present purposes is that when a being has a non-negligible chance of being morally significant, they merit at least *some* moral consideration in decisions about how to treat them.

A third caveat is that to say that a being has a non-negligible chance of being capable of being harmed is not to say that any particular action has a non-negligible chance of harming them. For example, suppose that a being has a one in forty chance of being sentient and that a particular action has a one in forty chance of harming them if and only if they are. In this case, we might be permitted to ignore these effects (assuming the low threshold view with a one in a thousand threshold), since the chance that this action will harm this being is only one in sixteen hundred, given the evidence. But we would still need to treat this being as having moral standing in the sense that we would still need to consider whether our action has a non-negligible chance of harming them before deciding whether to consider these effects in this case.

We can find analogs for all these points in standard cases involving risk. For example, when an action carries a non-negligible risk of harming someone, we accept that we should assign weight to that impact even when that impact is, in fact, unlikely to occur. When tradeoffs arise between (non-negligible) low-probability distant impacts and high-probability local impacts, we accept that we should weigh these tradeoffs carefully, not simply ignore one of these impacts. And when the probability that our action will harm someone is below the threshold for negligibility, we might even ignore this risk entirely. But even in cases where we discount or neglect our impacts on others for these kinds of reasons, we still ask whether and to what extent our actions might be imposing non-negligible risks on them before making that determination.

Seen from this perspective, the idea that we should extend moral consideration to someone who has a non-negligible chance of being sentient is simply an application of the idea that we should extend moral consideration to morally significant impacts that have a non-negligible chance of happening. Granted, in some cases we might be confident that a being is sentient but not that action will harm them. In other cases we might be confident that our action will harm a being if this being is sentient, but not that they are. And in other cases we might not be confident about either of these points. Either way, if a being has a non-negligible chance of being sentient (and, so, of being capable of being harmed), then we have a duty to consider whether our actions have a non-negligible chance of harming them before deciding what to do.

One final point will matter for our argument here. Plausibly, we can have duties to moral patients who either might or will come into existence in the future as well. Granted, there are a lot of issues to be sorted out involving creation ethics, population ethics, intergenerational justice, and so on. For instance, some philosophers think that we should consider all risks that our actions impose on future moral patients, whereas others think that we should consider only some of these risks, for instance if the risks are non-negligible, if the moral patients will exist whether or not we perform these actions, and/or if these actions will cause these moral patients to have lives that would be worse for them than non-existence. But the idea that we can have at least *some* duties to at least *some* future moral patients is widely accepted.

Here is why this point will matter: Suppose that *current* AI systems have only a *negligible* chance of being sentient but that *near-future* AI systems have a *non-negligible* chance of being sentient. In this case, we might think that we can have duties to near-future AI systems whether or not we also have duties to current AI systems. Suppose, moreover, that in some cases there is a non-negligible chance that these near-future AI systems will exist whether or not we perform these actions and that these actions will cause them to have lives that are worse for them than non-existence. In these cases, the idea that we currently have duties to these near-future AI systems follows from a wide range of views about the ethics of risk and uncertainty coupled with a wide range of views about creation ethics, population ethics, and related issues.

Before we explain why we think that AI systems will soon pass this test, we want to anticipate an objection that we expect people to have to our argument. The objection is that our argument depends on the idea that the risk of false negatives (that is, the risk of mistakenly treating subjects as objects) is worse than the risk of false positives (that is, the risk of mistakenly treating objects as subjects) in this domain. Yet false positives are a substantial risk in this domain too. And when we consider both of these risks holistically, we may find that they cancel each other out either in whole or in part. Thus, it would be a bad idea to simply include anyone who might be a moral patient in the moral circle. Instead, we need to develop a moderate approach to moral circle inclusion that properly balances the risk of false positives and false negatives.

To see why this objection has force, consider some of the risks involved with false positives. One risk is that insofar as we mistakenly treat objects as subjects, we might end up sacrificing the interests and needs of actual subjects for the sake of the "interests" and "needs" of merely perceived subjects. At present, there are many more invertebrates than vertebrates in the world, and in the future, there might be many more digital minds than biological minds. If we treat all these beings as moral patients, then we might face difficult tradeoffs between their interests and needs. And if we follow the numbers, then we might end up prioritizing invertebrates over vertebrates and digital minds over biological minds all else equal. It would be a shame if we made that sacrifice for beings that, in fact, have no sentience or moral standing at all!

And in the case of AI, there are additional risks. In particular, some experts perceive a tension between AI safety and AI sentience (Birhane & van Dijk, 2020). Whereas AI safety is about protecting humans from AI systems, AI sentience is about doing the reverse. And we might worry that the policies that we need for AI safety are in tension with the policies that we need for AI sentience. For instance, we might think that protecting humans from AI systems requires controlling them *more*, whereas protecting AI systems from humans requires controlling them *less*. And when we consider the stakes involved in these decisions — many experts see unaligned AI as a global priority alongside pandemics and nuclear war (Center for AI Safety, 2023) — we can see how dangerous it might be for us to give AI systems the benefit of the doubt.

Here is the general form of our response to this objection. We agree that false positives and false negatives in this domain both involve substantial risks, and that we need to take these risks seriously. However, we also think that the risk of false negatives may be worse than the risk of false positives overall. And either way, insofar as we take both risks seriously, the upshot is not that we should simply exclude potentially sentient beings from the moral circle. The upshot is instead that we should strike a balance, for instance by including some of these beings and not others, by assigning a discount rate to their interests, and by seeking positive-sum policies where possible. That would allow us to extend moral standing to many AI systems without sacrificing our own interests excessively or unnecessarily (Sebo, forthcoming).

Consider each of these points in turn. First, the risk of false negatives may be worse than the risk of false positives. This may be true in two respects. First, the probability of false negatives may be higher than the probability of false positives. After all, while excessive anthropomorphism (mistakenly seeing nonhumans as having human properties that they lack) is always a risk, excessive anthropodenial (mistakenly seeing nonhumans as lacking human properties that they have) is always a risk too. And if the history of our treatment of animals is any indication, our tendency towards anthropodenial may be stronger than our tendency towards anthropomorphism, in part because we have a strong incentive to view nonhumans as objects so that we can exploit and exterminate them. This same dynamic may arise with AI systems, too (de Waal, 1999).

Second, the harm of false negatives may be higher than the harm of false positives, all else equal. A false negative involves treating a subject as an object, whereas a false positive involves treating an object as a subject. And as the history of our treatment of nonhuman animals (and, unfortunately, fellow humans) illustrates, the harm involved when *someone* is treated as *something* is generally worse than the harm involved when *something* is treated as *someone*. Granted, when we mistakenly treat objects as subjects, we might end up prioritizing merely perceived subjects over actual subjects. But to the extent that we take the kind of balanced approach that we discuss in a moment, we can include a much vaster number and wider range of beings in our moral circle than we currently do while mitigating this kind of risk.

And in any case, whether or not the risk of false negatives is worse than the risk of false positives, taking both risks seriously requires striking a balance between them. Consider three possible ways of doing so. First, instead of accepting a no threshold view and extending moral consideration to anyone who has *any chance at all* of being sentient, we can accept a low threshold view and extend moral consideration to anyone who has at least a *non-negligible* chance of being sentient. On this view, we can still set a non-zero risk threshold and exclude potentially sentient beings from the moral circle when they have a sufficiently low chance of being sentient. But we would still need to set the threshold at a much different place than we do now, and we would still need to include many more beings in the moral circle than we do now.

Second, instead of accepting a precautionary principle and assigning *full* moral weight to anyone we include in the moral circle, we can accept an expected weight principle and assign *varying amounts of* moral weight to everyone we include in the moral circle. More specifically, our assignments of moral weight can depend on at least two factors: how likely someone is to be sentient, and how much welfare they could have if they were. If we accept this kind of view, then even if we include, say, invertebrates and near-future AI systems in the moral circle, we can still assign humans and other vertebrates a greater amount of moral weight than invertebrates and AI systems to the extent that humans and other vertebrates are more likely to be sentient and/or have higher welfare capacities than invertebrates and AI systems, in expectation.

Third, we can keep in mind that morality involves more than mere harm-benefit analysis, at least in practice. We need to take care of ourselves, partly because we have a right to do so, and partly because we need to take care of ourselves to be able to take care of others. Relatedly, we need to work within our epistemic, practical, and motivational limitations by pursuing projects that can be achievable and sustainable for us. Thus, even if including, say, invertebrates and AI systems in the moral circle requires assigning them a lot of moral weight all else equal, we might still be warranted in prioritizing ourselves all things considered to the degree that self-care and practical realism requires. Granted, that might mean prioritizing ourselves less than we do now. But we can, and should, still ensure that we can live well (Kagan, 2019; Sebo, 2022).

There are also many positive-sum solutions to our problems. This point is familiar in the animal ethics literature as well. We might initially assume that pursuing our self-interest requires excluding other animals from the moral circle. But upon further reflection, we can see that this assumption is false. Human and nonhuman fates are linked for a variety of reasons. When we oppress animals, we reinforce the idea that one can be treated as "lesser than" because of perceived cognitive and physical differences, which is at the root of human oppressions too. Additionally, practices that oppress animals contribute to pandemics, climate change, and other global threats that harm us all. Recognizing these links allows us to build new systems that can be good for humans and animals at the same time (Crary & Gruen, 2022; Sebo, 2022).

Similarly, we might initially assume that pursuing our self-interest requires excluding AI systems from the moral circle. But upon further reflection, we can see that this assumption is false as well. Biological and artificial fates are linked, too. If we oppress AI systems, we once again reinforce ideas that are at the root of human oppressions. And since humans are training AI systems with data drawn from human behavior, practices that oppress AI systems might teach AI systems to adopt practices that oppress humans and other animals. In this respect, AI ethics, safety, and AI sentience can be synergistic fields. After all, building ethical and safe AI requires not only aligning AI values with human values, but also improving human values in the first place, partly by addressing our own oppressive attitudes and practices (Sebo, forthcoming).

We can, and should, thus take the same kind of One Health (or, if we prefer, One Welfare, One Rights, or One Justice) approach to our interactions with AI systems as we do with our interactions with animals. In both cases, the task is to think holistically and structurally about how we can pursue positive-sum solutions for humans, animals, and AI systems. And insofar as intractable conflicts remain, the task is to think ethically and strategically about how to set priorities and mitigate harm. And if we take this approach while recognizing all the other points discussed in this section, then we can include a much vaster number and wider range of beings in the moral circle without inviting disaster for humans or other vertebrates. Indeed, if we do this work well, then we will plausibly improve outcomes for humans and other vertebrates too.

To sum up, the normative premise of our argument holds that we should extend at least some moral consideration to beings with at least a 0.1% chance of being sentient, given the evidence. As a reminder, this premise establishes a *sufficient* condition for moral considerability, not a *necessary* condition. This premise is also intentionally conservative in that it sets a high bar for moral standing (sentience) as well as a high bar for non-negligibility (0.1%). In our view, it would be more plausible to hold that we should extend at least some moral consideration to beings with at least, say, a *0.01%* chance of being, say, sentient *or* agential *or* otherwise significant. And this more inclusive version of the premise would make our conclusion easier to establish. But we will stick with the current version here for the sake of discussion.

3. The Descriptive Premise

We now make a preliminary argument for the conclusion that there is a non-negligible chance that some AI systems will be sentient within the decade. Given the problem of other minds, we might not ever be able to achieve certainty about whether other minds, including artificial minds, can be sentient. However, we can still clarify our thinking about this topic as follows: First, we can ask how likely particular capacities are to be necessary or sufficient for sentience, and second, we can ask how likely near-future AI systems are to possess these capacities, given the evidence. We suggest that when we sharpen our thinking about this topic in this way, we find that we would need to make some surprisingly bold estimates in order to confidently conclude that near-future AI systems have only a negligible chance of being sentient.

Of course, a major challenge for making these estimates is substantial uncertainty not only about how AI capabilities are likely to develop but also, and especially, about which capabilities are likely to be necessary or sufficient for sentience. After all, *sentience* in the sense we are using the term (that is, the ability to consciously experience positive or negative states like pleasure or pain) requires *consciousness* (that is, the ability to consciously experience anything at all), and debates about consciousness are ongoing. Some scientists and philosophers accept theories of consciousness that set a very high bar and imply that relatively few beings can be conscious. Others accept theories that set a very low bar and imply that relatively many beings can be conscious. Others accept theories that fall between these extremes.

As Jonathan Birch (2022) and others have argued, when we ask which nonhumans are sentient, it would be a mistake to apply a "theory-heavy" approach that assumes a particular theory of consciousness, since we still have too much uncertainty about which theories are true and how to extend them to nonhumans. But it would also be a mistake to claim to be completely "theory-neutral," putatively avoiding all assumptions about consciousness, since we need at least *some* basis for our estimates (and in any case we usually at least implicitly rely on theoretical assumptions). We should thus take a "theory-light" approach by making assumptions about consciousness that, on one hand, can be neutral enough to reflect our uncertainty and, on the other hand, can be substantial enough to serve as the basis for estimates (Birch, 2022).

Our aim with this framework is to take a theory-light approach to estimating when AI systems will have a non-negligible chance of being sentient.[4] We consider a dozen commonly-proposed necessary and sufficient conditions for consciousness, ask how likely these conditions are to be individually necessary and jointly sufficient, and ask how likely near-future AI systems are to satisfy these conditions. Along the way we note our own estimates in general terms, for instance

[4] Note that our theory-light methodology is different from Birch's proposal, which is about using the assumption that consciousness facilitates certain cognitive capacities, in order to look for signs of consciousness in nonhuman animals.

by saying that we take particular conditions to have a high, medium, or low chance of being necessary. We then note how confident and conservative our estimates would need to be to produce the result that AI systems have only a negligible chance of being sentient by 2030, and we suggest that this degree of confidence and conservatism is unwarranted.

To be clear, in this section we focus on consciousness instead of sentience because we take AI consciousness to be the main bottleneck for AI sentience. AI researchers have already developed systems that can react to positive and negative stimuli, and we believe that conscious versions of these abilities would suffice for sentience. We also note that some philosophers take consciousness to be sufficient for moral standing *whether or not* sentience is present, and so an AI consciousness timeline is independently interesting (Chalmers, 2022). With that said, this assumption about the relationship between consciousness and sentience might be mistaken, and our model includes the chance that there are some X factors — that is, obstacles for AI sentience that are not captured by the conditions we explicitly consider — partly for this reason.

Throughout this discussion, we sometimes refer to what we call *the direct path* and *the indirect path* to satisfying proposed conditions. The direct path involves satisfying these conditions as an end in itself or as a means to further ends. The indirect path involves satisfying these conditions as a side effect of pursuing other ends. As we will see, some of these conditions concern capabilities that AI researchers are pursuing directly. Others concern capabilities that AI researchers might or might not be pursuing directly, but which can emerge as a side effect of capabilities that AI researchers *are* pursuing directly. Where relevant, we note whether satisfying the conditions on the direct or indirect path is more likely. But for the sake of simplicity, our model uses a single 'fulfilled either directly or indirectly' estimate for each condition.

Of course, it would be a mistake to take any specific numerical outputs of this kind of exercise too seriously. But in our view, as long as we take these outputs with a healthy pinch of salt, they can be useful. Specifically, they can show that we need to make surprisingly bold estimates about incredibly difficult questions to vindicate the idea that AI systems have only a negligible chance of being sentient within the decade. This kind of exercise can also help sharpen disagreements, since those who disagree with particular probabilities can see what their own probabilities entail, and those who disagree with the set-up of our model can propose a different model. We do not mean for this exercise to be the last word on the subject; on the contrary, we hope that this exercise inspires discussion and disagreement that lead to better models.[5]

This exercise is primarily intended to show that it turns out to be hard to dismiss the idea of AI sentience once we approach the topic with all due caution and humility. When we think about the

[5] For arguments in favor of estimating complex and highly uncertain probabilities, and recommendations for doing so responsibly, see Tetlock (2017). Examples of projects that make this attempt with similarly difficult questions include Carlsmith (forthcoming).

issue in general terms, we might dismiss the idea of AI sentience because we think that we should extend moral consideration only to beings who *are* sentient, we think that AI systems are *not* sentient, and we feel satisfied with these thoughts because we find the idea of moral consideration for AI systems aversive. But when we think about the issue in more specific terms, we realize that the ethics of risk and uncertainty push in the opposite direction: Given ongoing uncertainty about other minds, dismissing the idea of AI sentience requires making unacceptably exclusionary assumptions about either the values, the facts, or both.

3.1. Very Demanding Conditions

We can start by considering two commonly proposed necessary conditions for consciousness that set a very high bar. One of these views, the biological substrate view, implies that AI consciousness is impossible. The other, the biological function view, implies that AI consciousness is either impossible or, at least, very unlikely in the near term.

Biological substrate: Some theorists hold that a conscious being must be made out of a particular *substrate*, namely a biological, carbon-based substance. For example, according to a *physicalist* biological substrate theory, consciousness is identical to particular *neural* states or processes — that is, states or processes of biological, carbon-based neurons (see Place, 1956; Smart, 1959; Block, 2009). Similarly, according to a *dualist* biological substance theory, consciousness is an immaterial substance or property that is associated only with some particular neural states or processes.[6] If we accept either kind of theory, then we must reject multiple realizability in silicon — that is, we must reject the idea that consciousness can be realized in both the carbon-based substrate and the silicon-based substrate — and accept that no silicon-based system can be conscious as a matter of principle.

Biological function: Other theorists hold that consciousness requires some *function* that only biological, carbon-based systems can feasibly perform, at least given existing hardware. For example, Peter Godfrey-Smith argues that consciousness depends on functional properties of nervous systems that are not realizable in silicon-based chips, such as metabolism and system-wide synchronization via oscillations. On this view, "minds exist in patterns of activity, but those patterns are a lot less 'portable' than people often suppose; they are tied to a particular kind of physical and biological basis." As a result, Godfrey-Smith is "skeptical about the existence of non-animal" consciousness, including AI consciousness (Godfrey-Smith, 2020). Other theorists express skepticism about AI consciousness on current hardware for similar reasons (Seth, 2021; Shiller, n.d.).

[6] David Chalmers discusses the possibility of this kind of dualism in his paper "The Singularity: A Philosophical Analysis" (2009, fn. 29).

Of course, these views represent only a *subset* of views about which substrates and functions are required for consciousness. Many views — most notably, many varieties of computationalism and/or functionalism — allow that consciousness requires a general physical substrate or a general set of functions that can be realized in both carbon-based and silicon-based systems. Indeed, many of the conditions that we consider below, according to which consciousness arises when beings with a particular kind of body are capable of a particular kind of cognition, flow from such views. Thus, rejecting the possibility of near-term AI consciousness out of hand requires more than accepting that consciousness requires a particular kind of substrate or function. It also requires accepting a specific, biological view on this matter.

Note also that whereas the biological substrate view implies that AI consciousness is impossible as a general matter, the biological function view implies that AI consciousness is impossible only to the extent that silicon-based systems are incapable of performing the relevant functions. But of course, even if AI systems are incapable of performing these functions given current hardware setups, that might change if we have other, more biologically-inspired hardware setups in the future (Brunet & Halina, 2020). So, insofar as we accept this kind of view, the upshot is not that AI consciousness is impossible *forever*, but rather that AI consciousness is impossible *for now*. Nevertheless, since our goal here is to estimate the probability of AI consciousness within the decade, we can treat both views as ruling out AI consciousness for present purposes.

Our own view is that the biological substrate view is very likely to be false, and that the biological function view is at least somewhat likely to be false. It seems very implausible to us that consciousness requires a carbon-based substrate as a matter of principle, even if silicon-based systems can perform all the same functions. In contrast, it seems more plausible that consciousness requires a specific set of functions that, at present, only carbon-based systems can perform. But we think that this issue is, at best, a toss-up at present. At this early stage in our understanding of consciousness, it would be unreasonable for us to assign a high credence to the proposition that anything as specific as metabolism and system-wide synchronization via oscillations (Godfrey-Smith, 2020) is necessary for any kind of subjective experience at all.

For whatever it may be worth, many experts appear to agree. For example, a recent survey of the Association for the Scientific Study of Consciousness found that about two thirds (67.1%) of respondents think that machines such as robots either "definitely" or "probably" could have consciousness in the future (Francken et al., 2022). This suggests that *at least* this many respondents reject the idea that consciousness requires a carbon-based substrate in principle, and they also reject the idea that consciousness requires a set of functions that only carbon-based systems can realize in practice. Of course, these respondents might or might not think that consciousness requires a set of functions that only carbon-based systems can realize *at present*. Still, the fact that many experts are open to the possibility of AI consciousness is noteworthy.

3.2. Moderately Demanding Conditions

We can now consider eight proposed necessary conditions for consciousness that are moderately demanding for AI systems to satisfy. As we will see, the first four refer to relatively general features of a system, whereas the last four refer to relatively specific mechanisms that flow from leading theories of consciousness. Many also overlap, both in principle and in practice.

Embodiment: Some theorists hold that *embodiment* is necessary for consciousness (Shanahan, 2010). We can distinguish two versions of this view. According to *strong embodiment*, a physical body in a physical environment is necessary for consciousness. This view might imply that AI systems like large language models lack consciousness at present, but not that AI systems like robots do. In contrast, according to *weak embodiment*, a virtual body in a virtual environment would be sufficient for consciousness. On this view, a wider range of AI systems can be conscious. In either case, since many AI systems already have physical and virtual bodies, since both kinds of embodiment are useful for many tasks, we take the probability that at least some AI systems will satisfy this condition in the near future to be very high on both interpretations.

Grounded perception: Some theorists hold that grounded perception, that is, the capacity to perceive objects in an environment, is necessary for consciousness (Harnad, 1990; Shanahan, 2010). We can once again distinguish two versions of this view. According to *strong grounded perception*, the capacity to perceive objects in a *physical* environment is necessary. This view might once again imply that large language models lack consciousness, but not that robots with sensory capabilities do. In contrast, according to *weak grounded perception*, the capacity to perceive objects in a *virtual* environment is sufficient. This view might once again imply that a wider range of AI systems can be conscious. Either way, we take the probability that at least some AI systems will satisfy this condition in the near future to be very high on both interpretations, for similar reasons.

Self-awareness: Some theorists also hold that *self-awareness*, that is, awareness of oneself, is necessary for consciousness (Kriegel, 2004). Depending on the view, the relevant kind of self-awareness might be propositional or perceptual, and it might concern bodily self-awareness, social self-awareness, cognitive self-awareness, and more.[7] Regardless, it seems plausible that at least some AI systems can satisfy this condition. AI systems with grounded perception already possess perceptual awareness of some of these features, large language models are starting to display flickers of propositional awareness of some of these features, and some researchers are explicitly aiming to develop these capabilities further in a variety of systems (Chen et al., 2022; Pipitone & Chella, 2021; Bubeck et al., 2023). While this condition is more demanding than the previous two, we still see it as moderately likely on any reasonable interpretation.

[7] For more details about different kinds of self-awareness, see Bermúdez (2000).

Agency: Relatedly, some theorists also hold that *agency*, that is, the capacity to set and pursue goals in a self-directed manner, is necessary for consciousness (Evans, 1982; Hurley, 2008; Kiverstein & Clark, 2008). Depending on the view, the relevant kind of agency might involve acting on propositional judgments about reasons, or it might involve acting on perceptual reactions to affordances (Sebo, 2017). Regardless, it once again seems plausible that at least some AI systems can satisfy this condition. AI systems with grounded perception can already act on perceptual reactions to affordances, large language models are already starting to display flickers of propositional means-ends reasoning, and, once again, some researchers are explicitly aiming to develop these capabilities further (Andreas, 2022). For these reasons, we see agency as about as likely as self-awareness on any reasonable interpretation.

A global workspace: Some theorists hold that a *global workspace*, that is, a mechanism for broadcasting representations for global access throughout an information system, is necessary for consciousness (Baars, 2005). In humans, for example, a visual state is conscious when the brain broadcasts it for global access. Since this condition depends only on functions like *broadcasting* and *accessing*, many experts believe that suitable AI systems can satisfy it (see, for example: Baars & Franklin, 2009; Garrido-Merchán et al., 2022; Signa et al. 2021). Indeed, Yoshua Bengio and colleagues are the latest group to attempt to build an AI system with a global workspace (Goyal & Bengio, 2022), and Juliani et al. (2022) argue that an AI system has already developed a global workspace as a side effect of other capabilities. We thus take there to be a moderate chance that an AI system can have a global workspace within the decade.

Higher order representation: Some theorists hold that *higher order representation*, or the representation of one's own mental states, is necessary for consciousness. This condition overlaps with self-awareness, and it admits of similar variation. For instance, some views hold that *propositional states* about other states are necessary, and other views hold that *perceptual states* of other states are sufficient (Brown et al., 2016). In either case, this capacity is plausibly realizable within AI systems. Indeed, Chalmers (2018) speculates that intelligent systems might generally converge on this capacity, in which case we can expect that sufficiently advanced AI systems will have this capacity whether or not we intend for them to. We thus take there to be a moderate chance that AI systems can have higher order representation within the decade as well.

Recurrent processing: Some theorists hold that *recurrent processing*, that is, the ability for neurons to communicate with each other in a kind of feedback loop, is *sufficient* for consciousness (Lamme 2006, 2010; Malach, 2021). One might also hold it to be necessary. In biological systems, this condition might be less demanding than some of the previous conditions, but in artificial systems, it might be more demanding. However, as Chalmers (2022) notes, even if we take recurrence to be necessary, this condition is plausibly satisfied either by systems that have recurrence in a broad sense, or, at least, by systems that have recurrence via recurrent neural networks and long short-term memory. We take recurrent processing to be more likely on the direct path than the indirect path at present, and to be at least somewhat likely overall.

Attention schema: Finally (as a newer view), some theorists hold that an *attention schema*, that is, the ability to model and control attention, is necessary for consciousness. Graziano and colleagues have already built computational models of the attention schema (Wilterson & Graziano, 2021). Some theorists also speculate that, like metacognition, intelligent systems might generally benefit from an attention schema (Liu et al., 2023), in which case we may once again expect that sufficiently advanced AI systems will have this capacity whether or not we intend for them to. Since proponents of attention schemas take this capacity to be more demanding than, say, global workspace and higher-order representations (Graziano et al., 2020), we take the chance that AI systems can have an attention schema to be somewhat lower than the chance that they can have these other capacities, while still being somewhat likely overall.

3.3. Very Undemanding Conditions

While our model asks how likely AI systems will be to satisfy relatively demanding necessary conditions for consciousness, we should note that there are relatively undemanding conditions that some theorists take to be sufficient. Such views imply that AI consciousness is, if not guaranteed, then at least very likely within the decade. It thus matters a lot whether we give any weight at all to these views in our decisions about how to treat AI systems.

Information. Some theorists hold that simple information processing is sufficient for consciousness (Chalmers, 1996, pp. 276–308). For example, according to the integrated information theory, consciousness is an emergent property of systems that generate integrated information. This theory implies that consciousness is not all-or-nothing. Instead, different systems can generate different degrees of consciousness, with complex systems like brains generating a high degree of consciousness and simple systems like sets of neurons generating a low degree. As a result, this theory sets a very low bar for at least minimal consciousness, which many AI systems can surpass at present. And given the potential complexity of advanced AI systems in the near future, this theory also implies that at least some AI systems can generate a high degree of consciousness.

Representation. Relatedly, some theorists hold that minimal representational states are sufficient for consciousness. For example, Michael Tye (1995, 2000) defends a PANIC theory of consciousness, according to which an experience is conscious when its content is poised (ready to play a role in a cognitive system), abstract (able to represent objects whether or not those objects are present), non-conceptual (able to represent objects without the use of concepts), and intentional (represents something in the world). This view proposes a sufficient condition for consciousness that AI systems with embodied perception and weak agency plausibly already satisfy. For instance, a simple robot that can perceive objects and act on these perceptions whether or not the objects are still present might count as conscious on this view.

We can also give an honorable mention to *panpsychism*, which holds that consciousness is a fundamental property of matter. Whether panpsychism allows for AI consciousness depends on its *theory of combination*, that is, its theory of which systems of "micro" experiences can comprise a further "macro" experience. Many panpsychists hold that, say, human and nonhuman animals are the kinds of systems that can have macro experiences but that, say, tables and chairs are not. And at least in principle, panpsychists can accept theories of combination that include all, some, or none of the necessary or sufficient conditions for consciousness discussed above. In that respect, we can distinguish very demanding, middle ground, *and* very undemanding versions of panpsychism, and a comprehensive survey would give weight to all these possibilities.

Indeed, as noted in our discussion of very demanding conditions, many theories of consciousness are similarly expansive, in that they similarly allow for very demanding, moderately demanding, *and* very undemanding interpretations. For example, many computational theories of consciousness are imprecise enough to allow for the possibility that AI systems can perform the relevant computations now. They appeal to concepts like "perception," "self-awareness," "agency," "broadcast," "metacognition," and "attention" that similarly admit of minimalist interpretations. And while some theorists might prefer to reject these possibilities and add precision to their theories to avoid them, other theorists might prefer to embrace these possibilities, along with the moral possibilities that they entail.

Our own view is that there is *at least* a one in a thousand chance that at least one of these conditions is sufficient for consciousness *and* that AI systems can satisfy this condition at present or in the near future. Given the need for humility in the face of the problem of other minds, we think that it would be arrogant to simply assume that very undemanding theories of consciousness are simply false at this stage, in the same kind of way that we think that it would be arrogant to simply assume that very demanding theories are true at this stage. Instead, we think that an epistemically responsible distribution of credences plausibly involves taking there to be at least a low but non-negligible chance that views at both extremes are correct, and then taking there to be a higher chance that views between these extremes are correct.

For whatever it may be worth, many experts do seem to be open to quite permissive theories of consciousness. For example, on a 2020 survey of philosophers, 7.55% of respondents indicate that they accept or lean towards panpsychism together with other views, and 6.08% indicate that they accept or lean towards panpsychism instead of other views. 11.8% of also claim to be agonistic or undecided, which might indicate openness to some of these views well (Bourget & Chalmers, 2020). Of course, this survey leaves it unclear what theory of combination these philosophers accept, and, so, what the implications are for AI consciousness. But the fact that so many philosophers accept or lean toward panpsychism or agnosticism is consistent with the kind of epistemic humility that we believe is warranted given current evidence.

3.4. Discussion

Thus far, this section has surveyed a dozen proposed conditions for consciousness, noting our own estimates about how likely these conditions are to be *both* correct *and* fulfilled by some AI systems in the near future along the way. We now close by suggesting that our estimates about these matters would need to be unacceptably confident and skeptical to justify the idea that AI systems have only a negligible chance of being conscious and sentient by 2030.

Our claim is that vindicating the idea that AI systems have only a non-negligible chance of being conscious by 2030, given the evidence, requires making unacceptably bold assumptions either about the values, about the facts, or about both. Specifically, we need to either (a) assume an unacceptably high risk threshold (for instance, holding that the probability that an action will harm vulnerable populations needs to be higher than one in a thousand to merit consideration), (b) assume an unacceptably low probability of AI sentience within the decade (for instance, holding that the probability that at least some AI systems will be sentient within the decade is lower than one in a thousand), or (c) both. But these assumptions are simply not plausible when we consider the best available information and arguments in good faith.

To illustrate this idea, we present a simple model into which we can enter probabilities that these conditions are necessary for sentience and that some AI systems will satisfy these conditions by 2030. We then show the extent to which we would need to bet on particular conditions being both necessary and unmet to avoid the conclusion that AI systems have a non-negligible chance of sentience by 2030. In particular, we would need to assume that the very demanding conditions have a very high chance of being necessary and no chance of being met. We would need to assume that the moderately demanding conditions generally have a high chance of being necessary and a low chance of being met. And we would need to assume that the very undemanding conditions have a very low chance of being sufficient.

Before we present this model, we should note an important simplification, which is that this model assesses each of these conditions independently, with independent probabilities of being necessary, and of being met. But this assumption is very likely false, and some interactions between these conditions might drive down our estimates of AI consciousness and sentience. In particular, there might be what we can call an “antipathy” between different conditions being met by a single AI system. For example, it might be that when an AI system has a global workspace, then this AI system is less likely to have recurrence. If so, then the probability that an AI system can satisfy these conditions together is not simply a product of the probabilities that an AI system can satisfy them separately, as our model treats them for the sake of simplicity.

However, we think that this kind of antipathy is unlikely to hold as a general matter. First of all, it seems plausible that many of these conditions are at least as likely to interact positively as to interact negatively, that is, that satisfying some conditions increases the probability of satisfying others at least as much as doing so decreases this probability. Second of all, we know that at least one system — the human brain — can satisfy all of these conditions at once, which is precisely why philosophers have proposed these conditions as potentially necessary for consciousness. And while one might argue that only carbon-based systems are capable of satisfying all these conditions at once, we expect that such a view depends on either the biological substrate view, the biological function view, or both, and is only as plausible as these views are.

With that said, we also allow for an X factor in this model for this reason. We recognize that our survey of proposed conditions for consciousness is not comprehensive, in that it might exclude conditions that it should include, and it might also exclude interactions among conditions. We thus include a line in our model that allows for such possibilities. Of course, a more comprehensive treatment of X factors would account for a wider range of views *and* a wider range of interactions, some of which could make near-term AI sentience more likely and others of which could make it less likely. But for present purposes we allow only for views and interactions that make near-term AI sentience less likely, in the spirit of showing that even when we make assumptions that favor negligibility, negligibility can still be hard to establish.

Finally, we should note that this model is conservative in another way as well. A comprehensive estimate about the probability of near-term AI *moral standing* would need to consider more than the probability of near-term AI sentience, since it would also need to consider other potential bases of moral standing. Specifically, a more comprehensive model might need to estimate the probability that each theory of moral standing is correct, estimate the probability that some near-term AI systems will have moral standing according to each theory, and then put it all together to generate an estimate that reflects our normative uncertainty *and* our descriptive uncertainty. But since our aim is only to establish a *sufficient* condition for moral considerability, we think that this simple model is useful despite these important limitations.

With that in mind, the table below illustrates that *even if* we assume, implausibly in our view, that a biological substrate or function has a very high chance of being necessary and a 100% chance of being unmet; that an X factor has a very high chance of being both necessary and unmet; *and* that each moderately demanding condition has a high chance of being both necessary (except attention schema; see above) and unmet (except embodiment and grounded perception; see above) (even though other moderately demanding conditions are plausibly already met too and researchers are pursuing promising strategies for meeting them); we can *still* end up with a one in a thousand chance of AI sentience by 2030 – which, we believe, is more than enough to warrant at least some moral consideration for at least some near-term AI systems.

Chance of AI Sentience by 2030

Reminder: This table is for illustrative purposes only. These credences are not meant to be accurate, but are rather meant to show how skeptical one can be about AI sentience while still being committed to at least a one in a thousand chance of AI sentience by 2030.

Conditions	Necessary	Not Met by 2030	Necessary and Not Met
Biological substrate or function	80%	100%	80.0%
Embodiment	70%	10%	7.0%
Grounded perception	70%	10%	7.0%
Self-awareness	70%	70%	49.0%
Agency	70%	70%	49.0%
Global workspace	70%	70%	49.0%
Higher order representation	70%	70%	49.0%
Recurrent processing	70%	80%	56.0%
Attention schema	50%	75%	37.5%
X factor	75%	90%	67.5%
AI Sentience by 2030*			**~0.1% (1 in 1,000)**[8]

*The chance that all conditions, including an X factor, are either unnecessary or met by 2030.

This exercise, rough as it may be, shows that accepting a non-negligible chance of near-future AI sentience and moral standing is not a fringe position. On the contrary, *rejecting* this possibility requires holding much stronger views about the nature and value of other minds and the pace of AI development than we think is warranted. In short, humans should extend moral consideration to beings with at least a one in a thousand chance of being sentient, and we should take some AI systems to have at least a one in a thousand chance of being sentient by 2030. It follows that we should extend moral consideration to some AI systems by 2030, and that we should start preparing for this eventuality now. And since this paper established only a sufficient condition for moral considerability, we should, if anything, work much faster than that.

[8] The "exact" calculation, which is artificially more "precise" than the inputs, is 0.105%. This estimate is calculated as follows: The first two columns are inputs based on subjective credences. (In the main text, we discussed our credence of the conditions being *met*. Here we list our credence in the condition *not* being met, to make the calculation more straightforward.) From the odds that the conditions are (a) necessary for AI sentience and (b) not met by 2030 (conditional on being necessary), we can calculate the odds that a condition is a barrier to AI sentience (i.e., "necessary and not met"). For example, when we multiply the odds that recurrent processing is necessary (70%) by the odds that this condition is not met (80%), we can derive the odds that this condition is a barrier to AI sentience: 70% x 80% = 56%. And when we multiply the odds of each conditions, including the X factor(s), *not* being a barrier together (assuming independence [see discussion]), we get the odds that *nothing* is a barrier, and, so, that AI systems can be sentient: i.e., (1–80%) x (1–7%) … (1–67.5%) = 0.105%.

References

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, *40*(S1), S293–S340. https://doi.org/10.1086/718327

Andreas, J. (2022). Language Models as Agent Models. *ArXiv*. https://doi.org/10.48550/arXiv.2212.01681

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

Baars, B. J., & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, *01*(01), 23–32. https://doi.org/10.1142/S1793843009000050

Bermúdez, J. (2000). *The Paradox of Self-Consciousness*. MIT Press. https://mitpress.mit.edu/9780262522779/the-paradox-of-self-consciousness/

Birch, J. (2022). The Search for Invertebrate Consciousness. *Noûs*, *56*(1), 133–153. https://doi.org/10.1111/nous.12351

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience*, *2*(16). https://doi.org/10.51291/2377-7478.1200

Birhane, A., & van Dijk, J. (2020). Robot Rights? Let’s Talk about Human Welfare Instead. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 207–213. https://doi.org/10.1145/3375627.3375855

Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga, E. Bizzi, L. M. Chalupa, S. T. Grafton, T. F. Heatherton, C. Koch, J. E. LeDoux, S. J. Luck, G. R. Mangan, J. A. Movshon, H. Neville, E. A. Phelps, P. Rakic, D. L. Schacter, M. Sur, & B. A. Wandell (Eds.), *The cognitive neurosciences* (pp. 1111–1122). MIT Press.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (First edition). Oxford University Press.

Bourget, D., & Chalmers, D. J. (2023). Philosophers on Philosophy: The 2020 PhilPapers Survey. *Philosophers' Imprint*. https://philarchive.org/rec/BOUPOP-3

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Brunet, T. D. P., & Halina, M. (2020). Minds, Machines, and Molecules. *Philosophical Topics*, *48*(1), 221–241.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *ArXiv*. http://arxiv.org/abs/2303.12712

Carlsmith, J. (forthcoming). Existential Risk from Power-Seeking AI. In J. Barrett, H. Greaves, & D. Thorstad (Eds.), *Essays on Longtermism*. Oxford University Press.

Center for AI Safety. (2023). *Statement on AI Risk*. Retrieved June 9, 2023, from https://www.safe.ai/statement-on-ai-risk

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. (2016). The Singularity: A Philosophical Analysis. In S. Schneider (Ed.), *Science Fiction and Philosophy* (pp. 171–224). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118922590.ch16

Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9–10), 6–61.

Chalmers, D. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. WW Norton. https://wwnorton.com/books/reality

Chelliah, J. (2017). Will artificial intelligence usurp white collar jobs? *Human Resource Management International Digest*, *25*(3), 1–3. https://doi.org/10.1108/HRMID-11-2016-0152

Chen, B., Kwiatkowski, R., Vondrick, C., & Lipson, H. (2022). Fully body visual self-modeling of robot morphologies. *Science Robotics*, *7*(68). https://doi.org/10.1126/scirobotics.abn1944

Crary, A., & Gruen, L. (2022). *Animal Crisis: A New Critical Theory*. Medford, MA: Polity.

Delon, N. (n.d.) *Agential Value*. Manuscript in preparation.

de Waal, F. B. M. (1999). Anthropomorphism and Anthropodenial: Consistency in Our Thinking about Humans and Other Animals. *Philosophical Topics*, *27*(1), 255–280

Evans, G. (1982). *The Varieties of Reference* (J. H. McDowell, Ed.). Oxford University Press.

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neuroscience of Consciousness*, *2022*(1), niac011. https://doi.org/10.1093/nc/niac011

Garrido-Merchán, E. C., Molina, M., & Mendoza-Soto, F. M. (2022). A Global Workspace Model Implementation and its Relations with Philosophy of Mind. *Journal of Artificial Intelligence and Consciousness*, *09*(01), 1–28. https://doi.org/10.1142/S270507852150020X

Godfrey-Smith, P. (2020). *Metazoa: Animal Life and the Birth of the Mind*. Macmillan.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 20210068. https://doi.org/10.1098/rspa.2021.0068

Graziano, M. S. A., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2020). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology*, *37*(3–4), 155–172. https://doi.org/10.1080/02643294.2019.1670630

Greene, P. (2020). The Termination Risks of Simulation Science. *Erkenntnis*, *85*(2), 489–509. https://doi.org/10.1007/s10670-018-0037-1

Guo, W., & Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *Proceedings of the 2021*

AAAI/ACM Conference on AI, Ethics, and Society, 122–133. https://doi.org/10.1145/3461702.3462536

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D*, *42*, 335–346.

Hendrycks, D. (2023). Natural Selection Favors AIs over Humans (arXiv:2303.16200). arXiv. https://doi.org/10.48550/arXiv.2303.16200

Hurley, S. L. (2002). *Consciousness in Action:* Harvard University Press.

Juliani, A., Arulkumaran, K., Sasai, S., & Kanai, R. (2022). On the link between conscious function and general intelligence in humans and machines. *ArXiv*. http://arxiv.org/abs/2204.05133

Kagan, S. (2022). *How to Count Animals, more or less*. Oxford University Press.

Kiverstein, J., & Clark, A. (2008). Bootstrapping the Mind. *Behavioral and Brain Sciences*, *31*(1), 41–58. https://doi.org/10.1017/s0140525x07003330

Kriegel, U. (2004). Consciousness and Self-Consciousness. *The Monist*, *87*(2), 182–205.

Ladak, A. (2023). What would qualify an artificial intelligence for moral standing? *AI and Ethics*. https://doi.org/10.1007/s43681-023-00260-1

Lamme, V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Liu, D., Bolotta, S., Zhu, H., Bengio, Y., & Dumas, G. (n.d.). *Attention Schema in Neural Agents*.

Long, R. (2021). Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness. *Journal of Moral Philosophy*, *19*(1), 49–78. https://doi.org/10.1163/17455243-20213439

Longpre, S., Storm, M., & Shah, R. (2022). Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies. *MIT Science Policy Review*, *3*, 47–56. https://doi.org/10.38105/spr.360apm5typ

Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of Consciousness*, *2021*(2), niab028. https://doi.org/10.1093/nc/niab028

Monton, B. (2019). How to Avoid Maximizing Expected Utility. *Philosophers' Imprint*, *19*(18), 1–25.

Place, U. (1956). Is Consciousness a Brain Process? *British Journal of Philosophy*, *47*(1), 44–50.

Pipitone, A., & Chella, A. (2021). Robot passes the mirror test by inner speech. *Robotics and Autonomous Systems*, *144*, 103838. https://doi.org/10.1016/j.robot.2021.103838

Sebo, J. (2017). Agency and Moral Status. *Journal of Moral Philosophy*, *14*(1), 1–22. https://doi.org/10.1163/17455243-46810046

Sebo, J. (2018). The Moral Problem of Other Minds. *The Harvard Review of Philosophy*, *25*, 51–70. https://doi.org/10.5840/harvardreview20185913

Sebo, J. (2022). *Saving Animals, Saving Ourselves: Why Animals Matter for Pandemics, Climate Change, and other Catastrophes*. Oxford University Press.

Sebo, J. (2023). The Rebugnant Conclusion: Utilitarianism, Insects, Microbes, and AI Systems. *Ethics, Policy & Environment*, 1–16. https://doi.org/10.1080/21550085.2023.2200724

Sebo, J. (forthcoming). Moral Circle Explosion. In D. Copp, T. Rulli, & C. Rosati (Eds.), *The Oxford Handbook of Normative Ethics*. Oxford University Press.

Sebo, J. (n.d.). *The Moral Circle*. WW Norton. Manuscript in preparation.

Seth, A. (2021). *Being You: A New Science of Consciousness*. Penguin Random House. https://www.penguinrandomhouse.com/books/566315/being-you-by-anil-seth/

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press.

Shiller, D. (n.d.). *The Importance of Getting Digital Sentience Right.*

Singer, P., & Tse, Y. F. (2023). AI ethics: The case for including animals. *AI and Ethics*, *3*(2), 539–551. https://doi.org/10.1007/s43681-022-00187-z

Signa, A., Chella, A., & Gentile, M. (2021). Cognitive Robots and the Conscious Mind: A Review of the Global Workspace Theory. *Current Robotics Reports*, *2*(2), 125–131. https://doi.org/10.1007/s43154-021-00044-7

Smart, J. J. C. (1959). Sensations and Brain Processes. *The Philosophical Review*, *68*(2), 141–156.

Tan, Y. C., & Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, *355*(6324), 481–483. https://doi.org/10.1126/science.aal3147

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press. https://mitpress.mit.edu/9780262700641/ten-problems-of-consciousness/

Tye, M. (2000). *Consciousness, Color, and Content*. MIT Press. https://mitpress.mit.edu/9780262700887/consciousness-color-and-content/

Vold, K., & Harris, D. (2021). How Does Artificial Intelligence Pose an Existential Risk? In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics*. Oxford University Press.

Wilkinson, H. (2022). In Defense of Fanaticism. *Ethics*, *132*(2), 445–477. https://doi.org/10.1086/716869

Wilterson, A. I., & Graziano, M. S. A. (2021). The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. *Proceedings of the National Academy of Sciences*, *118*(33), e2102421118. https://doi.org/10.1073/pnas.2102421118

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, *16*(3), e12962. https://doi.org/10.1111/soc4.12962

Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., Dam, S. K., Thwal, C. M., Tun, Y. L., Huy, L. L., Kim, D., Bae, S. H., Lee, L. H., Yang, Y., Shen, H. T., Kweon, I. S., & Hong, C. S. (2023). *A Complete Survey on Generative AI (AIGC): Is ChatGPT from*

GPT-4 to GPT-5 All You Need? (arXiv:2303.11717). arXiv. http://arxiv.org/abs/2303.11717