16

How Much Moral Status Could Artificial Intelligence Ever Achieve?

Walter Sinnott-Armstrong and Vincent Conitzer

Saudi Arabia recently granted citizenship to a robot.[1] The European Parliament is also drafting a form of "electronic personhood" for artificial intelligence.[2] Some Japanese get so attached to their robots that they give robots funerals and bury them after they break irreparably.[3] Many commentators see these recent developments as confused and even dangerous (Gunkel 2012), so we need to think about whether and why future artificial intelligence could or should ever be granted partial or even full moral status.

This chapter will begin by defining moral status and arguing that it comes in degrees on multiple dimensions. Next we will consider which conditions need to be met for an entity to have moral status, and we will argue that artificial intelligence can meet a combination of conditions that are sufficient for partial moral status. Finally, we will consider how much moral status an AI system could have.

1. What is Moral Status?

To understand the notion of moral status, consider common moral rules such as don't kill, don't disable, and don't deceive, among others. These rules seem simple, but they cannot be applied to the cases where moral status is at issue until we determine who it is that we should not kill, disable, or deceive. In short, which entities are protected by the moral rules? Another way of posing basically the same question is to ask whether an entity has moral rights, including the right not to be killed, disabled, or deceived. We can also ask whether other people have direct moral reasons not to treat the entity in certain ways or whether it is directly morally wrong to treat that entity in those ways. Asking about moral status is a shorthand way of asking which entities

Walter Sinnott-Armstrong and Vincent Conitzer, *How Much Moral Status Could Artificial Intelligence Ever Achieve?* In: *Rethinking Moral Status.* Edited by: Steve Clarke, Hazem Zohny, and Julian Savulescu, Oxford University Press.
 DOI: 10.1093/oso/9780192894076.003.0016

are directly protected by the four Rs: rules, rights, reasons, and wrongs (cf. DeGrazia 2008, 184).

Entities without moral status can still be protected *indirectly* by morality. It is morally wrong for someone else to blow up your car not because your car has moral status or rights but rather because you have moral rights not to have your property destroyed without permission, and blowing up your car will harm you. Your car is not wronged, but you are. In contrast, your pet dog has a right not to be burned alive, even if you want to commit that atrocity. That act wrongs your dog instead of wronging you, as in the case of your car. Thus, you and your dog are protected *directly* by morality insofar as what makes it wrong to harm you or your dog is something about you and your dog in contrast with anyone else who cares about you or your dog. That is what gives you and your dog moral status.

Of course, rules and rights can be violated justifiably, reasons can be overridden, and acts that are morally wrong in some circumstances can be justified in others. To say that an entity has moral status is not to say that is always immoral to kill, disable, or deceive it. It is only to say that it is directly morally wrong to kill, disable, or deceive it in situations where there is not enough reason to do so.

2. Does Moral Status Come in Degrees?

Some philosophers claim that each entity simply has moral status or not. One example is Elizabeth Harman, who says, "…moral status is not a matter of degree, but is rather on/off: a being either has moral status or lacks it" (Harman 2003, 183). Harman does admit that a human counts more than an anaconda, but only because death causes a greater loss to the human than to the anaconda. Regarding pain, for example, pain to the anaconda counts less than pain to the human, because the human will remember the pain longer, will suffer more while remembering it, and will have more projects that the pain prevents the human from accomplishing. Harman insists, nonetheless, that equal harms to different beings with moral status create equally strong moral reasons.

We disagree. To see why, imagine a human to whom the pain or other moral wrong means no more than to the anaconda. Perhaps the human will die very shortly after being harmed, so the human will have no memories or projects for the pain to interfere with. However we set up this example, there should be some way to ensure that the human will not lose significantly more

than the anaconda. Nonetheless, in a case where each entity loses the same amount, it still seems more morally wrong to harm the human than to harm the anaconda. Reflection on examples suggests that moral status comes in degrees.[4]

In particular, moral status varies in degree along (at least) two dimensions: strength and breadth. To see how moral rights can vary in strength, compare an anaconda, a bonobo, and a human child. If you could not save both the anaconda and the bonobo from death, or if you could not avoid killing one of them, then it would seem immoral to kill or fail to save the bonobo instead of the anaconda. But what if you could not save both the bonobo and the human child or could not avoid killing one of these? Then it seems (except to extremists on animal rights) immoral to kill or fail to save the human child instead of the bonobo. These comparisons suggest that the bonobo's moral right not to be killed is stronger than any such right in the anaconda but weaker than the human child's right.

Moral status also varies in breadth, that is, how many rules, rights, reasons, and wrongs protect a certain entity. For example, babies have rights not to be tortured or killed, but they have no rights not to be deprived of freedom. It is not immoral to swaddle them tightly even when their squirming suggests that they want to be free. But it would be immoral to do anything like this to any normal adult human, such as put them in a straightjacket (even for the adult's own good, if that is why we swaddle babies). Thus, babies have the same right not to be caused pain, but they do not have the same right to freedom as adults.

Conversely, imagine an otherwise normal adult human who cannot feel any pain because of an unchangeable biological deficit.[5] This permanently numb adult can still have moral rights to be free and not to be killed or disabled. However, it makes little sense to say that this permanently numb adult has a moral right not to be caused pain, because it is constitutionally unable to feel any pain. Opponents might reply that the numb adult has other properties that give it a conditional moral right not to be caused pain *if* it did somehow become able to feel pain. However, there might be no way for that ability to arise without changing the numb human's biology so much that it becomes a different organism and person. Moreover, a moral right conditional on other circumstances is not a moral right not to be caused pain now, while it cannot feel pain because of how it is currently constituted.[6]

These degrees of moral status are crucial here, because we will argue that a future AI with certain features can have a moral right to freedom but no moral right not to be caused pain, much like the numb adult or an angel,

according to some theologies. This conclusion is controversial, and we admit our own doubts. But before we can argue for it, we need to address one more preliminary issue.

3. What is the Basis of Moral Status?

It is not enough merely to announce *that* an entity has moral status. One must specify *why* it does. This reason is the basis for its moral status, rights, or protection.

The properties that supply this basis must meet certain standards to be fair, explanatory, and not question begging. We agree with Bostrom and Yudkowsky (2014), who argue for two limitations on which properties can be the basis for moral status. First:

> *Principle of Substrate Non-Discrimination*: If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status. (p. 323)

In short, what matters is not substrate but function. To see why, imagine that a doctor discovers that your best friend is actually Neanderthal rather than human. Would that make your friend's moral status questionable? No, despite genetic differences. Your friend's moral status would not be in doubt even if the doctor found that her body was made of silicon instead of carbon. What matters is her consciousness, intelligence, and other functions rather than their physical substrate. This point will become crucial when we come to the question of whether computers or AIs can have moral status.

Bostrom and Yudkowsky's (2014) second principle concerns source or origin:

> *Principle of Ontogeny Non-Discrimination*: If two beings have the same functionality and the same conscious experience, and differ only in how they came into existence, then they have the same moral status. (p. 324)

Again, imagine that your best friend tells you that a mad scientist somehow created her from frog cells, using CRISPR to modify the genes. She would still be intelligent, conscious, and your friend, so she would have full moral status. Thus, origin does not matter to moral status any more than substrate.

Analogously, the fact that AIs come from programmers in a very different way than humans come from parents cannot show that they lack moral status.

Finally, a basis for moral status would be useless in determining which entities have moral status if it were not also empirically determinable (Liao, forthcoming). For example, a theory that AIs as well as fetuses and animals have moral status just in case they have souls cannot help us unless it also provides some way to tell which entities have souls. We need that help, so such theories are practically inadequate, even if they are theoretically defensible.

3.1 Sentience

One popular and plausible proposal for the basis of moral status is sentience, which is the capacity to experience feelings, sensations, emotions, or moods. In arguing for animal rights, DeGrazia (this volume) prominently claims that sentience is necessary and sufficient for moral status. What matters to the issue of animal rights is sufficiency, but what matters regarding AI is necessity. If sentience is necessary for moral status, and if AIs are not sentient, then AIs cannot have moral status.

We doubt that sentience is necessary for all moral rights or status. It would be necessary for a moral right not to be caused pain, since a non-sentient creature cannot feel pain. However, it is not at all clear that or why sentience would be necessary for a moral right to life, freedom, privacy, or speech, since sentience is not necessary for life, freedom, privacy, or speech.

To see this point, imagine that a human is prevented from achieving his goals but feels no pain or even frustration, perhaps because he does not know that he failed to achieve what he wanted. His right to freedom still might be violated. Similarly, if the camera on his laptop secretly records him, this violates his right to privacy, even if he never finds out and never experiences any consequences of having been recorded. Again, his right to speech is violated if the government blocks his email (a form of speech) without him ever discovering that his protest messages never got through. Even his right to life can be violated by killing him painlessly in his sleep so that he is never aware of being killed and never feels any pain or frustration. Because such victims' rights can be violated without any negative feelings that their sentience makes them able to sense, it is hard to see why sentience would be necessary for a right to freedom or those other rights.

Imagine also that we encounter sophisticated aliens who are not sentient at all. A tenuous peace between them and us emerges, and we all manage to get along and work together towards our objectives. It would violate their rights if we broke our promises to them or killed or enslaved them. Granting such basic rights to them seems essential to maintaining our peaceful arrangement with them. If so, at least some creatures without sentience can have some rights, so sentience is not necessary for all moral rights.

The same goes for interests if interests require felt desire or felt frustration when those interests are not met (*pace* DeGrazia, this volume). In contrast, if interests are merely goals that shape an entity's behavior, then they might be relevant to freedom, because we cannot restrict the freedom of entities to pursue their goals if they have no goals. But then there is no reason why an advanced AI in the far future could not have goals that shape its behavior, so it could have interests of this kind and then moral rights.

DeGrazia might reply that plants have this kind of biological goals and interests, but plants do not have a right to freedom, so how can interests be sufficient for a right to freedom? The solution is either to distinguish the kinds of interests that plants have from the kinds that ground moral status or to hold that goals convey moral status only in the context of intelligence, agency, and other abilities that plants lack.[7] We do not and need not claim that interests by themselves are sufficient for a moral right to freedom.

3.2 Multiple bases

The fundamental problem with requiring sentience or felt interests for any moral status is that they are relevant to some moral rights (such as the right not to be caused pain), but they are irrelevant to other moral rights (such as rights to freedom and life). A theory of moral status is better when it cites properties that explain not only *which entities* have at least some moral status but also *which rights* they have, which is to say *how broad* their moral status is. It is doubtful that any single property can explain such different rights.

It seems preferable to align different features of the affected entity as the basis of the different rights, reasons, rules, and wrongs that apply to that entity. The right not to be caused pain seems to require sentience, whereas the right to be free seems to require goals together with the ability to make rational choices. Neither of these requirements depends on substrate or ontogeny, and both are empirically determinable, so they meet the main

requirements for bases of moral status, even if they do not provide a unified basis for all kinds of moral status.

4. Can Future AIs have the Basis of Moral Status?

We can now answer the question of whether an advanced AI far in the future could meet the conditions for moral status. As we saw, AIs cannot be excluded from moral protection either because they lack cells or were programmed or because they lack felt interests or sentience (interpreted narrowly as the capacity to experience feelings, sensations, emotions, or moods). But plants show that merely having unfelt goals is not sufficient for a moral right to freedom (and, hence, for some degree of moral status) in the absence of other abilities.

Which abilities? Plausible and popular candidates include intelligence, consciousness, freedom, and perhaps also moral understanding. We will not try to determine which of these abilities is individually necessary for a moral right to freedom. What matters here is they are jointly sufficient, and nothing else (including sentience) is necessary.

We will argue that an advanced AI far in the future could have all of these abilities. Since they are jointly sufficient for moral status, showing that AIs can have them all will be enough to show that an advanced AI far in the future could have this much moral status.

4.1 Intelligence

The name "AI" means artificial *intelligence*, but we should not infer too much from this name. The fact that something is called artificial intelligence does not show that it is really intelligent. Similarly, people often say that an air conditioning system is *trying* to cool down the house, so they attribute intentions, but they don't *really* believe that the system has desires or a model of what it is trying to achieve, much less a concept of the house. This common and useful way of speaking and thinking does not show that air conditioners are really intentional or intelligent.

A reverse mistake is to think that a system is not intelligent just because we know how it works. Computer science students are sometimes assigned to write a simple program for playing a simple game, such as connect-four

(in which players take turns dropping discs from the top and try to get four discs of their color in a row). Even very talented students who play against their own algorithms and understand exactly how those algorithms work find it too difficult to think through all the moves ahead that the algorithm considers. It is more effective for them to imagine that they are playing against another human, and then they end up thinking that the algorithm is *trying* to achieve certain goals. But we still understand exactly what is going on, at least in principle: a systematic but rote enumeration of all relevant moves. And with this understanding of how the connect-four AI works, it is natural to say that it isn't really intelligent. This is known as the "AI effect": once AI researchers figure out how to accomplish a benchmark task, observers dismiss the achievement by saying, "Well, but that's not really intelligence." That assessment would be unfair. The accomplishment certainly tells us *something* about the nature of intelligence and seems to display *some* kind of intelligence, even though it is hard to verbalize exactly which kind.

Probably in part due to this difficulty, the AI community has not agreed on a single definition of intelligence. The most popular definitions are pragmatic, flexible, and inclusive. A very inclusive definition might say only that intelligence is any ability to acquire and apply knowledge and skills. This definition seems to capture one common meaning, and AI seems able to acquire and apply knowledge and skills. AI can acquire knowledge or information, for example, simply by searching the internet for data. It can apply that knowledge in reaching conclusions, such as predictions about what people will buy or how they will vote. It can acquire skills, such as how to play games, and it can apply that skill by beating humans.

Max Tegmark seems to require more when he defines intelligence as "the ability to accomplish complex goals" (Tegmark 2017, p. 39). Does an AI that beats humans at chess really have winning as its goal? Maybe winning is a goal for the programmer but not for the AI itself. However, one sign that an entity has a goal is that the goal guides its actions. When our goal is to win rather than just to play or have fun, we will adjust our moves in ways that increase the probability of winning, even if those moves make the game less fun and shorter. That is exactly what an AI does when it plays chess. A learning AI may even adjust the weights of its connections so that it will become more likely to win next time. Even though it got this goal from its programmer, and regardless of whether it is conscious of this goal, it is guided by the goal of winning. Thus, AI can fit Tegmark's definition of intelligence as well.

A much less inclusive way of defining and testing intelligence was proposed by Turing (1950). In the Turing test, a player sends and receives messages

from two sources, one human and one computer. The computer is supposed to display intelligence to the extent that the player cannot tell them apart. If an AI ever passes this Turing test,[8] that achievement is supposed to show that the AI has intelligence.

But is the Turing test the right standard for intelligence? One problem with the Turing test is that it requires general intelligence about all topics that the player might ask about. We do not see why this much range is required, since an entity can have intelligence without having all kinds of intelligence. A savant who can quickly calculate the day of the week for every day in the past century is displaying unusual intelligence on that topic, even if his intelligence is very limited on other topics. So passing the Turing test should not be seen as necessary for intelligence.

Is it sufficient? Critics claim that the computer is only simulating intelligence without having any real intelligence. The best-known and most forceful argument for this objection is probably Searle's Chinese room thought experiment (1980). Searle asks us to imagine a person inside a room with no access to the outside except Chinese characters that others send in occasionally. The person does not read Chinese but has a large instruction manual that tells him which Chinese characters to put out when certain Chinese characters come in. The person does not understand either the characters or what he is doing. Searle argues that understanding is necessary for real intelligence, computers are analogous to the person in the Chinese room, and their programs are analogous to the translation manual, so AI cannot really have understanding or intelligence.

This argument had force against the kinds of computers and programming that existed at the time when Searle introduced his argument. However, the analogy arguably breaks down with the recent progress observed in machine learning. AI that uses machine learning can develop new skills that were not programmed into it. A programmer who is a relatively poor Go player could program a computer to beat the world Go champion at the game of Go. The AI achieves success by playing itself millions of times and changing its strategy in accordance with its wins and losses. Changing its strategy can be seen as a way of reprogramming itself.[9]

This kind of learning makes the AI very different from the instruction manual or the person in Searle's Chinese room, since those never learn or change in order to better meet their goals. They couldn't do that without knowing their goals and also knowing when those goals are met, which requires more (and more varied) access to the outside than merely receiving inputs occasionally. And when we add these other elements (especially the

ability to rewrite the translation manual to achieve known goals), then it is not at all clear why we should not see the person as understanding and as intelligent. In principle, we could also see the system or the room *as a whole* as learning through notes that the manual instructs the human to write down on paper and periodically consult. In this case, the human may neither learn nor understand anything, and the same is true for the manual, but arguably the room as a whole is doing both.

The point is that advanced AI methods make computers or programs more closely analogous to human intelligence. In deep learning, artificial networks resemble (to some degree) what happens in our brains when we learn. These methods have been remarkably effective at achieving complex goals. And just as it is hard for one human to figure out what is going on in another human's brain, it is also generally quite difficult to assess what exactly is happening in these artificial networks.[10] Their achievements, arguable similarity to human brains, and opacity incline people to see such deep learning networks as thinking. Indeed, Geoffrey Hinton, who has been playing a major role in the deep learning revolution, is not shy about ascribing "thoughts" to artificial neural networks.[11]

It is becoming clear, however, that these networks, at least for now, are not doing *exactly* the same thing as our brains. One difference is shown by the susceptibility of such networks to so-called adversarial examples. Even when algorithms correctly label most images, changing a few pixels in an image often results in the algorithm completely mislabeling what to us are completely unambiguous images. The algorithm picks up on some statistical pattern in the data it has seen, but often the pattern is more about local texture than about a complete understanding of the image. In contrast, humans interpret images in light of more global contexts, so they are rarely fooled by such minuscule changes.

It is good to remain aware that such algorithms can sometimes obtain impressive performance without much, if any, thinking or understanding. Consider the example of finding your way through a corn maze. This problem might seem to require a good amount of intelligence, including keeping track of where you are, whether you have been here before, and what you've already explored. However, a simple trick that works for many mazes is (spoiler alert!) simply to continue to follow the wall on your left-hand side. If you didn't know this simple trick, and an AI system discovered it, then you might imagine that the AI models the maze and its own place in the world. That would seem very intelligent and impressive. In reality, however, it is not doing anything like that. It is just following the left-hand wall, for which it doesn't

even need to remember anything. And its achievement is no more impressive just because no human can figure out how it is doing so well.

The main message is this. When humans use certain cognitive capacities to perform well on a task, and then an AI system performs as well or better on that same task, this still does not mean that that the AI has the same cognitive capacities as the human. Sometimes an algorithm does solve a problem in a similar way as we do, but it can be difficult to know when it does, especially in the case of deep learning.

In the context of this chapter, which argues that different capacities imply different moral rights, it is therefore crucial not to confuse tasks with capacities. Good performance on a task does not necessarily mean that the system has the underlying capacity that a human uses for the task. Then again, when we believe that intelligence is what is required for a particular task or right, it is not clear why that intelligence must work in the same way as our human intelligence. Moreover, even if we cannot be certain whether a particular AI is intelligent, it is still possible that some advanced AI far in the future could somehow come to possess our level of intelligence or more. Then it will become hard to say why we have moral rights based on our intelligence, but that AI does not have similar moral rights based on its intelligence.

Many more objections could be raised. Nonetheless, we conclude tentatively that advanced AI far in the future can have any kind of intelligence that is required for moral rights and status.

4.2 Consciousness

Another property that is often said to be necessary or sufficient for moral status is consciousness. Although this claim is common, it is not at all clear which kind of consciousness is supposed to determine moral status.

One crucial distinction is between phenomenal and access consciousness (Block 1997). *Access* consciousness is merely access to information. An entity has access consciousness of an orange when it can see it, grab it, or count it when asked how much fruit is there. It lacks access consciousness of the orange when it does not detect the orange and cannot form beliefs or make decisions in light of the information that the orange is there.

Phenomenal consciousness is more mysterious. An entity has phenomenal consciousness when there is something that it is like to be that entity or have that entity's experiences. A human who has been completely color blind since birth does not have phenomenal consciousness of the color orange and does

not know what it is like for a human with color vision to see the color orange. Nonetheless, color blind humans can still get access to information about which objects are orange by asking other people, so they can have access consciousness of the color orange without phenomenal consciousness of that color.

Which kind of consciousness matters to moral status? Our answer should not be surprising after the preceding discussion. Different kinds of consciousness matter to different moral rights that constitute different aspects of moral status. Phenomenal consciousness matters to the right not to be caused pain, because the way that pain feels is essential to what pain is.[12] Thus, an entity without any phenomenal consciousness of pain cannot feel pain and, hence, cannot have a right not to be caused pain. In contrast, access consciousness is crucial for rational decisions, which require access to information about one's options. An entity that cannot rationally decide to do or not do a certain act is not really free to do or not do that act, so it makes little sense to grant it a moral right to freedom. That is why babies have a moral right not to be caused pain but no moral right to freedom, because they have enough phenomenal consciousness to feel pain but not enough access consciousness to consider the information needed to make rational and free decisions.[13]

These distinctions enable a more fine-grained position on the moral status of AI. It is not clear *how* to begin to build phenomenal consciousness into AI. It is also not clear *why* anyone would do so. What good would it do? Pain is said to have evolved in biological organisms partly in order to detect tissue damage, but an AI could use other methods to detect damage to its parts. Another proposed evolutionary purpose of pain is to prevent organisms from moving injured parts in ways that might slow recovery or lead to re-injury (Klein 2015). But, again, AI could avoid such dangerous movements by using other sources of information about what not to move. AI would not need pain for these purposes, so they could not provide any reason to build AI so that it could feel pain. Humans might try to create an AI that feels pain in order to experiment on it and thereby learn more about pain, but the ethics of such experiments would be dubious if the AI really did feel pain. Moreover, even if programmers did program pain into an AI or if some advanced AI accidentally came to feel pain, it would still be difficult to tell whether an advanced AI really feels pain, much less the same kind of pain that we do. And even if our kind of pain requires phenomenal consciousness, it would remain questionable whether the AI has phenomenal consciousness until we better understand what this kind of consciousness is, what produces it, and how it affects behavior.

In any case, our main points here are conditional. If an AI cannot feel pain, it will have no right not to be caused pain. But even if an AI does not feel pain or experience any phenomenal consciousness, that is not enough to show that it does not have any moral rights, because it still might have moral rights that are unconnected to phenomenal consciousness, including, possibly, the right to freedom. An AI that does not feel pain could still access information and use it in making choices, seeking goals, and performing tasks. It would then have the kind of access consciousness that is needed for rational decisions.[14] That would be a basis for its moral right to freedom. Overall, then, an advanced AI far in the future might have moral status with regard to freedom but not with regard to pain.

Some critics might object (as Frances Kamm did in conversation) that phenomenal consciousness is (obviously?) also required for a moral right to freedom. This position seems plausible to many,[15] but it is not immediately clear how to formulate a strong argument for this requirement. Although phenomenal consciousness, including pain, affects people's choices, that does not mean that people cannot have goals and choose means to those goals in a rational way without phenomenal consciousness. That was the lesson from the numb adult described above. Restricting freedom is morally significant because it prevents agents from achieving their rational goals. This basis for the moral right to freedom does not require phenomenal consciousness. Access consciousness is enough, at least in some cases.

The point is not just that an AI with access consciousness but no phenomenal consciousness can have a derivative right to freedom. Suppose that Alice agrees to go on a lifelong deep-space mission under the condition that an AI system gets to pursue Alice's goals on Earth unimpeded. Maybe now the AI system has a right to freedom, but this right derives from Alice's rights. It is *Alice's* right that others not interfere with her AI system.

Our claim instead is that an AI with access consciousness but no phenomenal consciousness can have its own rights. To see how, imagine that a highly advanced vacuuming robot is noisily vacuuming the public space where you are currently taking a telephone call. You would like it to go vacuum somewhere else for a while, and promise it that you will finish your call and get out of its way after five minutes. The robot recognizes that it will be able to achieve its goals better if it agrees to your request, so it does. It might even point out to you that it does not have to grant your request, because the policy for the public space is that robots can vacuum anywhere at any time, so you should not break your promise. Does the robot now have a right that you finish and leave within five minutes? We think that it does, that there is no reason to

think that this right depends on its having phenomenal consciousness, and that its right does not derive solely from the rights of those managing the public space. This kind of example suggests that an AI can have certain abilities that are sufficient for certain moral rights even without phenomenal consciousness.

None of this is meant to deny that what is happening on the inside matters. It is important to emphasize this, because the AI research and development community has generally focused on external performance of systems. This community *does* care about what is happening on the inside, but usually *only* insofar as the inside affects external performance. In contrast, what is happening on the inside does matter independently of external performance when we are talking about whether AI has certain moral rights. Our thesis here is not that phenomenal consciousness never matters. It does matter sometimes. Our claim is only that not all direct or underived moral rights depend on phenomenal consciousness, so an AI can have some moral rights of its own even if it has no phenomenal consciousness.

4.3 Free will

A moral right to freedom might seem to require more than access consciousness to information needed for rational decisions. Something like free will might also seem necessary. After all, if free will is required for moral responsibility, as many assume,[16] and if an entity could have a moral right to freedom without having free will, then it would be morally wrong to restrict that entity's freedom while that entity would not be morally responsible for restricting the freedom of other moral agents. That seems unfair.

Instead of criticizing this line of reasoning, we will argue that an AI can have free will in any sense that matters. This claim depends, of course, on what free will is. Contemporary philosophers typically assume naturalism and deny that free will or moral responsibility requires any immaterial soul (Mele 2014; Nadelhoffer 2014) or any uncaused action (*pace* Kane 2007). But then what is necessary for free will? Philosophers disagree about the answer.

One of the most popular views is that agents act of their own free will when and only when their decisions and actions result from a mechanism that is responsive to reasons for and against those decisions and actions (Fischer 2007). To say that a mechanism is responsive to certain reasons is simply to say that the mechanism has access to information about the reasons and reacts appropriately, so that it (or the agent who uses that mechanism) does

an action when there is overriding reason to do it and does not do the action when there is overriding reason not to do it. If such responsiveness is enough for free will, then an advanced AI far in the future could have both. We already saw that AI can have access to information about reasons for and against decisions and actions, and it can adjust its behaviors to that information. Thus, AI can have reasons-responsiveness and free will, according to this theory.

Another popular theory proposes that agents act of their own free will when their actions mesh properly not only with their first-order desires to do those actions but also with their second-order desires to have those first-order desires (Frankfurt 1988). This theory implies that a drug addict who is happy to be an addict takes drugs freely, whereas an addict who regrets and fights against addiction does not take drugs freely. An advanced future AI could also meet these conditions for free will. If first-order desires are just dispositions to behave in certain ways, and if an AI can reprogram itself to change its dispositions so as to better achieve its goals, then we can understand its disposition to reprogram itself as a second-order desire to change its first order-desires. This structure is exactly what is required for free will, according to this theory.[17]

More generally, an advanced AI far in the future will be able to satisfy any conditions required by any plausible naturalistic theory of free will.[18] Opponents still might insist that reasons-responsiveness and higher-order mesh are not sufficient for *real* free will, perhaps because *real* free will requires a soul or uncaused actions, which AI cannot have. In response, we would deny that such non-naturalistic free will is necessary for moral responsibility. Reasons-responsiveness and higher-order mesh are enough for an AI to be morally responsible for restricting the freedom of the other moral agents, so it would be unfair not to admit that it would also be morally wrong to restrict that AI's freedom. What really matters here is moral responsibility rather than free will.

4.4 Moral understanding

Another common requirement on moral responsibility is the ability to understand or appreciate moral reasons, rights, rules, and wrongs. It is unfair to hold people responsible for doing something wrong when they could not have known that what they did was wrong. This requirement on responsibility has been assumed by most legal insanity defenses since the 1500s (Sinnott-Armstrong and Levy 2011).

Will any advanced AI ever be able to tell right from wrong? We saw in the preceding section that an AI can be responsive to reasons, and that responsiveness could extend to moral reasons. However, responsiveness to reasons is not yet enough to ensure understanding of those reasons as reasons. So we still need to ask whether any advanced AI could ever understand moral reasons.

How can we tell whether other humans (such as students) understand any proposition? One common method is to ask them to draw inferences from and give reasons for that proposition. The same standard holds in morality. When someone knows *that* she morally ought to keep her promises in general, knows *when* (in which cases) she morally ought to keep her promises, knows *why* she morally ought to keep her promises, and knows *what follows* for interpersonal relations and punishment from the proposition that she morally ought to keep her promises, then these abilities together are enough evidence that she understands the moral reasons for keeping promises.

These conditions on understanding could in principle be met by an advanced future AI. Our team at Duke is currently trying to program morality into a computer or AI (Freedman et al. 2020). Our method is to determine which features humans take to be morally relevant and how those features interact in order to produce human moral judgments. Our machine learning techniques then apply these human views about morally relevant features and their weights, so they should be interpretable, at least in principle. The resulting program should be able to predict *which* actions humans judge to be morally wrong or not and also able to specify *why* those actions are morally wrong or not by citing the very same features of those actions that humans themselves would give as reasons for their moral judgments. We might not be able to understand how all of these features interact or precisely how to define each feature, but we should at least be able to understand roughly which features play a role in the model, because those features came from surveys of humans. So far, our team is only in the initial stages of developing this method in a pilot study of kidney transplants. We have a long way to go. Still, if our method succeeds eventually, then the resulting artificial intelligence will be able to tell us *that* an act is wrong, *when* it is wrong, *why* it is wrong, and *what follows* from the fact that it is wrong. These are the abilities that show moral understanding in humans, so they will be enough to show that the resulting AI also has moral understanding. An AI with all of these abilities will understand the what, when, and why of moral reasons. Of course, an AI may not understand these as deeply as human beings do, perhaps because the AI

understands the world in general less well, but it is sufficient for it to have *some* understanding.

Critics might reply (as Frances Kamm did in conversation) that moral responsibility requires not only moral understanding but also phenomenal consciousness of moral wrongness. However, as we argued, it is not clear why phenomenal consciousness is required for all moral rights. Moreover, we do not even know or ask about agents' phenomenal consciousness of moral wrongness—what it is like for them to recognize an act as wrong[19]—before we hold them responsible. This leaves no barrier to our claim that AI in the far future might have the kind of moral understanding that is relevant to moral responsibility and rights.

5. Conclusion

Our overall argument is simple: An advanced AI far in the future could have (the relevant kind of) intelligence, (access) consciousness, (naturalistic) free will, and (functional) moral understanding. Anything with all of these properties can have some moral rights. Rights (along with reasons, rules, and wrongs) are all there is to moral status. Therefore, a future AI can have some degree of moral status.

We still do not know how much breadth of moral status an AI could have. We suggested that a future AI could have a right to some kinds of freedom (and against interference with such freedom), even if it has no right not to be caused pain. This is the opposite of a human baby, which has a right not to be caused pain (or at least tortured) but not a right to freedom, such as to move where it wants. The reason is that the baby can feel pain, whereas the AI cannot (we are assuming for now[20]); and the AI can access information to make rational choices, whereas the baby cannot. Their differing abilities and vulnerabilities determine their rights.

This simple contrast leaves a host of questions about other rights. What about:

A right to life and to defend itself?
A right to nutrition (electricity) and to health (parts)?
A right to speak or to associate?
A right to education or updating?
A right to procreate or to get married?
A right to vote or to serve on juries?

Some of these issues are hard to resolve, in part due to our lack of understanding of phenomenal consciousness. Despite these and many other open questions, our argument is enough to show that an AI can have moral status with some but not unlimited breadth, even if we do not know exactly how broad it is.

Notes

1. <https://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html>.
2. <https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues>.
3. <https://www.nationalgeographic.com/travel/destinations/asia/japan/in-japan—a-buddhist-funeral-service-for-robot-dogs/>.
4. This position has been held by Buchanan 2009, DeGrazia 2008, Persson (this volume), and others.
5. As Frances Kamm pointed out, another example might be an angel who can make rational decisions but has no body, so it cannot feel pain or be killed.
6. The numb adult retains a moral right that others not harm it by damaging its tissues even if it lacks a right not to be caused the pain that indicates or prevents that harm in normal humans (Klein 2015).
7. Brain parts also lack such abilities. Suppose Lok deliberates and decides not to have any dessert after dinner tonight, but some subconscious part of his brain makes thoughts pop up—how wonderful ice cream would taste, why it wouldn't be so bad to have a little, and how easy it would be to get some. These thoughts make his hand reach for the ice cream in an inattentive moment. This subconscious part of his brain might *seem* to have a goal and an intelligent method of achieving it all on its own. Does it (as opposed to Lok as a whole person) have a right to freedom? We don't think so, partly because Lok has a right to suppress it. Choice and agency apply only to a person or entity as a whole instead of its parts, so it is Lok rather than a lobe of his brain that has a right to freedom.
8. Despite some claims to have passed a Turing test (<https://www.bbc.com/news/technology-27762088>), we doubt that any computer today could pass a serious Turing test with plenty of time, knowledgeable judges, and human contestants who are motivated to win.
9. AI with this type of learning does not reprogram itself in the way humans program a computer. It does not write new code. What it does is adjust the weights of connections between nodes in its network, which changes the probabilities that activation of one node will affect others. AI systems come closer to what we normally think of as programming when they make use of meta-learning and architecture searches. Thanks to Nick Bostrom for this point.
10. Is this opacity the same as we saw in the connect-four program? Well, not quite. In the case of neural networks (unlike the connect-four program), it is often hard for the AI

programmers and researchers to get an accurate idea, even in the abstract, of what exactly the network is doing. Most of the time, their focus is on how well the network performs rather than how it performs so well.

11. <https://medium.com/syncedreview/geoffrey-hinton-on-images-words-thoughts-and-neural-patterns-82db0bd04a09>.
12. This point might be challenged by theories the understand pain in terms of preferences or imperatives (Klein 2015) instead of phenomenology, but they make phenomenal consciousness even less relevant to moral status.
13. Of course, we can sometimes be justified in restricting the freedom of people, such as teenagers, to prevent them from hurting themselves or others, but that shows only that their right to freedom can be overridden. Teenagers are still different from babies, because teenagers have rights that need to be overridden by strong reasons, whereas we can swaddle infants for very little reason or even no reason at all other than our own convenience or tradition. That suggests that babies have no right to freedom, whereas teenagers do.
14. The best current theory of access consciousness in biological organisms is the global workspace theory of Dehaene (2014). On that theory, consciousness emerges from neural loops that cause massive increases in brain activity. Exactly the same kind of wiring could be built into AI.
15. Including one of us (Conitzer).
16. This assumption is denied by semi-compatibilists (Fischer 2007; Vierkant et al. 2019), but we will not question it here.
17. Higher-order and mesh theories are often understood as kinds of deep-self theories (Sripada 2016). Critics might object better AI cannot have a self, much less a deep self. However, deep-self theories are naturalistic and do not require any metaphysically extravagant kind of self. All they require are cares, desires, or commitments of a kind that an advanced AI far in the future could have.
18. An innovative and plausible naturalistic account of free will has recently been developed recently by Christian List (2019), who explicitly says that strong AI could have free will on his account.
19. Indeed, we doubt that there is any unified phenomenology of moral judgments (Sinnott-Armstrong 2008).
20. If a future AI did somehow become able to feel pain, for whatever reason, then it might gain a right not to be caused pain. But that would give it more moral status instead of less.

References

Block, N. (1997). "On a Confusion about a Function of Consciousness." In *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Guzeldere. Cambridge, Mass.: MIT Press.

Bostrom, N., and Yudkowsky, E. (2014). "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, ed. K. Frankish and W. M. Ramsey, pp. 316–34. New York: Cambridge University Press.

Buchanan, A. (2009). "Moral Status and Human Enhancement." *Philosophy and Public Affairs* 37 (4): 346–81.

DeGrazia, D. (2008). "Moral Status as a Matter of Degree?" *Southern Journal of Philosophy* 46 (2): 181–98.

DeGrazia, D. (this volume). "An Interest-Based Model of Moral Status."

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts*. New York: Penguin.

Fischer, J. M. (2007). "Compatibilism." In *Four Views on Free Will*, by J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas, pp. 44–84. Malden: Blackwell.

Frankfurt, H. (1988). "Freedom of the Will and the Concept of a Person." Reprinted in H. Frankfurt, *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Freedman, R., Schaich Borg, J., Sinnott-Armstrong, W., Dickerson, J. D., and Conitzer, V. (2020). "Adapting a Kidney Exchange Algorithm to Align with Human Values." *Artificial Intelligence* 283: 103261. DOI:10.1016/j.artint.2020.103261.

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, Mass.: MIT Press.

Harman, E. (2003). "The Potentiality Problem." *Philosophical Studies* 114: 173–98.

Kane, R. (2007). "Libertarianism." In *Four Views on Free Will*, by J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas (pp. 5–44). Malden: Blackwell.

Klein, C. (2015). *What the Body Commands: The Imperative Theory of Pain*. Cambridge, Mass.: MIT Press.

Liao, S. M. (forthcoming). "The Moral Status and Rights of Artificial Intelligence." In *The Ethics of Artificial Intelligence*, ed. S. M. Liao. New York: Oxford University Press.

List, C. (2019). *Why Free Will is Real*. Cambridge, Mass.: Harvard University Press.

Mele, A. (2014). "Free Will and Substance Dualism: The Real Scientific Threat to Free Will?" In *Moral Psychology, Volume 4: Free Will and Moral Responsibility*, ed. W. Sinnott-Armstrong, pp. 195–208. Cambridge, Mass.: MIT Press.

Nadelhoffer, T. (2014). "Dualism, Libertarianism, and Scientific Skepticism about Free Will." In *Moral Psychology, Volume 4: Free Will and Moral Responsibility*, ed. W. Sinnott-Armstrong, pp. 209–16. Cambridge, Mass.: MIT Press.

Persson, I. (this volume). "Moral Status and Moral Significance."

Searle, J. (1980). "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3: 417–57.

Sinnott-Armstrong, W. (2008). "Is Moral Phenomenology Unified?" *Phenomenology and the Cognitive Sciences* 7(1): 85–97.

Sinnott-Armstrong, W., and Levy, K. (2011). "Insanity Defenses." In *The Oxford Handbook of Philosophy of Criminal Law*, ed. John Deigh and David Dolinko, pp. 299–334. New York: Oxford University Press.

Sripada, C. (2016). "Self-expression: A Deep Self Theory of Moral Responsibility." *Philosophical Studies* 173: 1203–32.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf.

Turing, A. M. (1950). "Computing Machinery and Intelligence." *Mind* 59 (236): 433–60.

Vierkant, R., Deutschländer, R., Sinnott-Armstrong, W., and Haynes, J.-D. (2019). "Responsibility without Freedom? Folk Judgements about Deliberate Actions." Frontiers in Psychology, Cognitive Science Section, 10, Article 1133.