Robots with Moral Status?

David DeGrazia

ABSTRACT Assuming robots of the future will be far more advanced than their present-day forebears, it is not premature to ask what they will have to be like in order to have moral status. This article first examines criteria for moral status, criticizing several models before briefly defending an interest-based account. It next investigates the epistemological challenge of applying criteria for moral status to robots, before eliciting implications with attention to basic moral status, rights, and respect for autonomy. The article concludes with reflections on species-based prejudice and an acute practical dilemma that will confront robotics.

Siri and Alexa, autonomous vehicles and weapons systems, robot diagnosticians and caregivers, and the ever-evolving internet represent a state of artificial intelligence (AI) that was unimaginable a generation ago. Looking ahead, we may anticipate an explosion of robotics technology, big data capabilities, deep machine learning, and other forms of AI far surpassing what exists today. The purpose of this article is to identify and reflect on some very important questions well before the state of technology demands answers of us. Although some of the

Department of Philosophy, George Washington University, Rome Hall 566, Washington, DC 20052.
Email: ddd@gwu.edu.

The author would like to express his gratitude for helpful feedback and encouragement to the editors of this journal, to an anonymous reviewer, and to several members of the George Washington University Department of Philosophy who commented on a draft.

Perspectives in Biology and Medicine, volume 65, number 1 (winter 2022): 73–88.

reflections apply to non-robotic AI systems, this discussion will focus on robots. By *robots* I mean programmable machines that interact with their environment using sensors and that can perform actions at least somewhat independently of their programmers. AI involves the development of computer programs that can perform tasks that would otherwise require human (or at least organic) intelligence. For reasons that will become clear, the robots of interest in this essay will instantiate AI.

Assuming robots of the future will be far more advanced than their present-day forebears, it is not premature to ask what they will have to be like in order to have *moral status*. What traits would robots need in order to matter morally in their own right and have rights or at least morally weighty interests? I claim that robots will gain moral status if and when they acquire their own *interests*—and, collectively, a welfare that matters to them—which will happen if (and only if) they become sentient. Moreover, in order to become sentient, robots will have to achieve consciousness, because sentience is the capacity for consciousness that features pleasant or unpleasant experiences. How we can know whether robots are conscious, however, is an enormous epistemological problem. In addition, there is a practical dilemma. Robotics is advancing with an eye toward serving human interests—for example, doing tedious chores for us, performing complex medical procedures, offering companionship, and engaging in dangerous military operations. Yet the very advances that make robots proficient at their assigned tasks might eventually transform them into sentient beings with interests of their own. And that eventuality would provoke legitimate concerns about exploitation and even slavery.

The issues just identified—the possible moral status of future robots (and other AI systems), the challenge of knowing whether they are conscious, and concerns about our wrongfully exploiting such artifacts if and when they exist—have been engaged in a fairly well-developed literature. So what can the present discussion add? The first intended contribution is to help bring these issues to the attention of bioethicists and others, including medical professionals, who work in the medical humanities. The developed literature on this topic appears to be consumed primarily by specialists in AI and AI ethics, with few of the articles published in more general ethics journals. (For a recent article that helps to bring epistemological issues to readers of bioethics and medical humanities, see Shevlin 2021). The second intended contribution is to engage these issues in a unique way. Although several of my main ideas are shared by others (as citations will indicate), the analyses of moral status and autonomy, the emphasis on the distinction between consciousness and sentience, a suggestion I advance about the point of moral status, and the commentary on parallels between robot ethics and animal ethics are my own.

The remaining discussion begins by exploring criteria for moral status, criticizing several models before suggesting, and briefly defending, an interest-based account. The next section investigates the epistemological challenge of applying criteria for moral status to robots. This section also elicits some implications of robots' moral status with attention to basic moral status, rights, and respect for autonomy. The article concludes with reflections on species-based prejudice and the aforementioned practical dilemma that will confront robotics. By the end, I hope to have shown how future developments in robotics will usher in a second battle (the first concerning nonhuman animals) between those who believe that species membership is central to our moral status and those who deny this traditional assumption.

Criteria for Moral Status

What criteria should we apply in considering whether robots might acquire moral status? Traditionally, there has been a tendency to assume that only members of our species, *Homo sapiens*, have moral status. But this assumption has been put in doubt—many would say refuted—in recent decades by leading work in animal ethics (see, for example, Armstrong and Botzler 2017; Beauchamp and Frey 2011).

That humans don't have a monopoly on moral status should be evident upon careful reflection. Tormenting a cat for enjoyment is morally wrong. A sufficient reason to judge such behavior wrong is that it harms the cat extensively and gratuitously. This judgment implies that the cat's interests have moral importance in their own right—and that the cat herself matters morally in her own right. In other words, the cat has moral status and is not merely a resource for human use or enjoyment.

Unpacking the concept of moral status more precisely might prove helpful as we consider possible bases for moral status. Consider this analysis: *X has moral status if and only if (1) moral agents have obligations regarding their treatment of X and (2) it is for X's sake that moral agents have these obligations.* This means that only beings who have a "sake"—a prudential or self-interested standpoint—can have moral status. I find it helpful to conceptualize the idea of a "sake" in terms of interests. So here is a revised analysis: *X has moral status if and only if (1) X has interests, (2) moral agents have obligations regarding their treatment of X, and (3) these obligations are responsive to X's interests* (see DeGrazia 2008).

Returning to our cat, the conditions of moral status are straightforwardly met: the cat has interests (for example, not to be hurt or harassed); we have certain obligations regarding our treatment of cats (for example, not to abuse them); and a clear basis for our obligations regarding the cat is the cat's interests and, more generally, her welfare. So we may be confident that moral status is not the exclusive domain of humans. Moreover, our reasoning seems to generalize to all sentient animals—since all have an interest in avoiding torment, or harm more generally, grounding an obligation not to harm them gratuitously. For this reason, it makes sense to regard sentience as a sufficient basis for moral status.[1]

Note that cats who are susceptible to tormenting and who are good candidates for moral status are not only sentient but (rather obviously) *alive*. So is life—being biologically alive—an important basis for moral status? I suggest that the best answer is no. (For works defending an affirmative answer, see Taylor 1986; Varner 1998; Warren 1997. For a work that engages the possible moral status of AI systems and suggests that being alive is one of several criteria bearing on moral status, see Liao 2020.)

Consider living things such as bacteria, fungi, plants, and very primitive animals such as sponges that lack consciousness and therefore sentience. Living things, it is often asserted, have biological needs. They need water, nutrition, and freedom from destruction in order to survive and reproduce. Notice, however, that organisms that lack the capacity for consciousness cannot *care* whether they survive and reproduce. While we might say they need certain things relative to these goals, that doesn't mean that they have real interests or any "sake" for which we may act.

Consider comparisons. A car needs oil and gasoline in order to run properly. A beautiful painting needs freedom from vandalism in order to remain beautiful. The moon needs not to be obliterated in order to continue to exist. But neither the car nor the painting nor the moon has interests, because none of them has any experiential relationship, positive or negative, to what happens to it. Such objects have no point of view, so to speak, and therefore no prudential standpoint—without which, interests are inconceivable. I suggest that, in the same way, living things such as plants that lack the capacity for consciousness, and therefore sentience, have no interests despite their biological needs. Their needs are relative to certain biological goals, but there is no meaningful sense in which *they* (not being subjects) have those or any other goals. Being alive, in my view, is insufficient for having interests, so it is also insufficient for moral status. Suppose, however, I am wrong about this, and plants have interests grounded in biological needs. Even then, as John Basl and Joseph Bowen (2020) argue, it is plausible to judge that plants' interests are not very weighty in comparison to those of sentient beings, not weighty enough to confer a substantial moral status (or, as they emphasize, rights).

[1] John Basl (2014) defends a somewhat similar position through a different argumentative route. For example, Basl argues that many nonsentient beings, including plants, have "teleo interests," but that the latter lack direct moral importance, whereas I argue that nonsentient beings lack interests. In addition, my discussion distinguishes basic moral status from rights possession and explores the importance of autonomy. In a more recent article, Basl and Joseph Bowen (2020) take an interest-based approach to the possession of moral rights, arguing that all and only conscious beings possess interests weighty enough to generate rights and, correlatively, obligations in others. By contrast, I distinguish moral status from rights possession and contend that consciousness is necessary, but not sufficient, for having interests and moral status. As will become clear, this latter thesis is important because some future robots might be conscious yet lack sentience.

Might being alive be *necessary* for moral status? After all, the nonhuman beings to whom we are inclined to attribute moral status—namely, certain animals (not their remains)—are all living. But maybe that observation rests on the contingent fact that so far there have been no strong nonliving candidates for moral status. That situation could change—and indeed, it will change if robots become sentient and acquire interests. What will matter is their possession of interests, not whether they are alive. (For readings that address criteria for moral status with an eye on robotics, see Bostrom and Yudkowsky 2014; Coeckelbergh 2010; Gunkel 2018; Levy 2009; Schwitzgebel and Garza 2015. For a recent book that focuses on animal and environmental law as a basis for defending robot rights, see Gellers 2021.)

Let me proceed to what I believe to be a more defensible account of moral status. (For a more fully developed argument in favor of this account, see DeGrazia and Millum 2021, chap. 7.) As our analysis of this concept suggests, only beings with interests have moral status. But do *all* beings who have interests have this status? Although I cannot prove an affirmative answer, I submit that a negative answer is unreasonable. How could some beings with interests of their own, and experiential welfares, count for nothing from the moral point of view? It would seem to involve a kind of bigotry to withhold all moral consideration from them. So I maintain that having interests is not only necessary, but also sufficient, for moral status.

Now we need to ask which beings have interests. A sensible answer is that all and only sentient beings do (see Singer 1975; Steinbock 2011). *Sentience* is the capacity to have pleasant or unpleasant experiences. Any being who can have such experiences has an experiential welfare: subjective experience that can go well or badly from the being's point of view. That is enough to confer interests, at least an interest in a good quality of life. Sentience is also necessary for having interests, I maintain, because a being or entity that is entirely incapable of having pleasant or unpleasant experiences cannot care what happens to it. Importantly, I construe the terms *pleasant* and *unpleasant* broadly, so that they apply not only to sensory experiences that are closely associated with hedonism—such as sensory pleasure and pain—but also to emotional states, such as satisfaction and frustration, and to any form of caring. Beings who have any such mental states are sentient and have interests and moral status, on my account. By contrast, beings who do not care about anything and find nothing pleasant or unpleasant, attractive or aversive, lack interests and moral status.[2]

Sentience, I have argued, is necessary and sufficient for moral status. Moreover, for purposes of discussion I will assume here that all beings with moral status have a substantial, rather than trivial, level of moral status (DeGrazia 1996). It's not as

[2] One might maintain that a conscious being who lacked sentience but had values might have interests and moral status—contrary to my claim that only sentient beings make these grades. I address this possibility later.

if canaries, being sentient, have moral status, but at a level so trivial that moral agents may override canaries' most important interests—for example, to fly—just to satisfy their own trivial desires, such as the desire to keep pretty birds in cages. Note, however, that to say that all beings with moral status have some sort of substantial moral status is compatible with holding that only some of them—say, persons—enjoy the additional moral protections of rights, where rights generally serve to protect an individual's most important interests from being sacrificed on utilitarian grounds. One view of this kind, which I find promising, holds that beings (namely, persons) with the sort of robust self-awareness involved in *narrative identities*—in which one conceptualizes one's own life as forming a sort of story—have special longer-term interests that ground the additional moral protection of rights (see DeGrazia and Millum 2021). But this is a controversial claim that will play a relatively minor role in this discussion. For our purposes, the most important thesis about moral status is that sentience is necessary and sufficient for (nontrivial) moral status.

Rather than proving this thesis about sentience, I have only motivated it with several arguments. Any readers who are not persuaded may profitably read the remainder of this article in a conditional way: *if* sentience is necessary and sufficient for (nontrivial) moral status, *then* this [what I go on to argue] is a sensible way to think about the possible moral status of future robots. Surely even this conditional claim is worth exploring, insofar as it elicits the implications of an important model of moral status. Others might wish to explore the implications of alternative models—which is all to the good, because we should think carefully and open-mindedly about the possibility of robots having moral status well before the issue confronts us with a feeling of urgency.[3]

How Can We Know If a Robot Has Moral Status?

On the present account, the possibility of future robots with moral status rests on whether they will have interests, which in turn will depend on whether they are sentient. There are different ways in which robots might be sentient. They will have *sensation-based* sentience if, say, they have a tactile sense and can experience mechanical, thermal, or chemical changes with a positive or negative feel—permitting pain, discomfort, or tactile pleasure. This seems much more likely in robots that can move around and touch things than in, say, the computer HAL in *2001*, who was immobile if very thoughtful (see Clarke 1968).

Robot sentience is also possible, in principle, in a purely *emotional* form. Imagine a robot that lacked sensory feelings but cared about accomplishing certain

[3] According to some models, moral status is grounded not only in an individual's properties, such as sentience or being alive, but also in one's relationships to other individuals (see Coeckelbergh 2010; Gellers 2021; Kittay 2005). I disagree, but space constraints preclude a proper rebuttal here. For discussions that cast substantial doubt on the relationship-based approach, see Jaworska and Tannenbaum 2018; Liao 2020.

aims. "My job is to get Junior to school on time and, dang it, this is the second time this week I got him there late!" Caring about achieving its aims would entail sentience, because the caring would typically generate satisfaction or frustration at the achievement or thwarting of aims.

Because sentience requires the capacity to *feel*, not simply information-processing (which present-day computers, robots, and other AI systems engage in), sentience is impossible without *consciousness*. By "consciousness" I mean nothing more than *subjective experience*, whatever its nature and however it may come about. (Some philosophers call this "phenomenal consciousness"; see Block 1995.) As I understand the *concept* of consciousness, it is not susceptible to further analysis, because this concept and those of subjectivity, experience, awareness, and what Nagel (1974) calls "what-it-is-like-to-be"-ness are equally basic and defy further analysis—except, circularly, in terms of each other.[4] One therefore might have to clarify these terms by "pointing" to consciousness—that is, by directing someone to consider what's always present in waking and dreaming states and absent in dreamless sleep or under general anesthesia.

Would Robot Consciousness Confer Moral Status?

Is it even possible for robots, presumably super-sophisticated robots, to be conscious? Some think not. These thinkers assume that consciousness can only be generated by, or realized in, certain kinds of physical material like carbon-based flesh, which would not be used to build robots. (Let us assume, for discussion's sake, that the robots under consideration are not "bio-bots" that incorporate neural tissue into a mostly robotic body. Presumably bio-bots could become conscious, once technical obstacles are overcome, because effective machine-brain interfaces already exist and bio-bots would have neural tissue, which we know can support consciousness.) Other thinkers believe consciousness requires an immaterial substance, or soul, which could not connect in the right way to an artifact. The skeptical positions, of course, beg the perennial question of how minds and bodies ultimately relate to each other.

I will assume it is possible, in principle, for robots to become conscious for the simple reason that we are in no position to preclude this possibility: for starters, we don't know enough about the mind to know which theory of the mind-body relation is correct. Thus, like Basl and Bowen (2020, 287), I practice "substrate nondiscrimination"—an open attitude about what sorts of materials might manifest or produce consciousness—while understanding that my working assumption might actually be false. If it is false, then robots will never become conscious and will always lack moral status. If my assumption is true, then robots might in fact

[4] This conceptual point is compatible with the idea that consciousness itself, the phenomenon, may admit of analysis. The concept of consciousness is distinct from the nature of consciousness, just as the concept of water is distinct from its empirically discovered essential structure, H_2O. For a classic discussion of the distinction in the case of water and other natural kinds, see Kripke 1980.

become conscious someday. Alternatively, if we repeal the previous paragraph's assumption and allow bio-bots to count as robots, then there is a far greater likelihood that conscious robots will emerge someday, regardless of which account of the mind-body relation is correct. Either way, we should get a running start in thinking about robots' moral status and how we should treat them. As Susan Schneider (2020) argues, we should be thinking about how to test for consciousness in machines lest we end up wronging conscious robots (or AI systems more generally) without even realizing they are entities who can be wronged. My friendly amendment to this thesis is that we should work on developing tests for sentience, not just consciousness, for reasons already explained. But consciousness is a precondition of sentience, so first things first.

Granting the in-principle possibility of robot consciousness, we face the daunting epistemological question of how we can know, or responsibly believe, whether a particular robot is conscious. To some extent, this question parallels the question of animal consciousness—how can we know, for example, whether crustaceans are conscious?—but in one major respect it differs. Unlike nonhuman animals, robots do not share an evolutionary lineage with *Homo sapiens*. So we can't appeal to neuroanatomical analogies as a type of evidence, since robots (other than bio-bots) don't have neuroanatomies. Nor can we appeal to evolutionary function, since robots didn't evolve through natural selection. The epistemological challenge—knowing whether a particular advanced robot is conscious—might seem insurmountable. Yet, ethically, we cannot evade this challenge once robots become sufficiently advanced that consciousness begins to seem possible—any more than we can ethically evade the challenge of judging whether the lobsters we boil alive, or the horseshoe crabs we use in research, have subjective experiences such as pain. (For a thoughtful discussion of the interplay between our uncertainty in the face of the epistemological challenge and the ethics of interacting with advanced forms of AI, see Agar 2020.)

Scholars have begun thinking about tests that might provide evidence of consciousness in robots or other AI systems. Two approaches, developed by Schneider and Edwin Turner (as discussed in Schneider 2020), strike me as especially promising. One, the Chip Test, would approach the question of machine consciousness somewhat indirectly, by exploring whether a particular type of inorganic material used in a microchip is capable of sustaining consciousness. The test might proceed in patients who—say, because they had brain tumors—were appropriate candidates for having parts of their neural tissue replaced by microchips designed to replicate relevant forms of brain functioning. Microchips would replace bits of a subject's brain gradually, with intervals in which subjects could report whether they experienced any changes suggesting loss of conscious function. If they did not, this result would suggest that the material used in the microchips was capable of sustaining conscious function (at the very least in a human host), in which case it would seem promising to use this material in constructing the "brain" of a robot designed to achieve consciousness.

The second approach, called the AI Consciousness Test, or *ACT* for short, involves interviewing a robot (or other AI system) whose sophisticated verbal behavior indicates a good candidate for a conscious subject. Can the robot convincingly answer questions that seem impossible to answer without direct familiarity with consciousness and some facility in imagining its presence or absence? Such questions might concern the imagined possibilities of a non-bodily afterlife, reincarnation, or an out-of-body experience. Does the robot seem to understand such puzzles as the "hard problem" concerning the nature of consciousness; "zombie" cases featuring unconscious beings that behave exactly as conscious ones would; or the idea of "inverted spectra," in which, say, A and B both use "red" and "green" in describing blood and grass respectively, though their subjective color experiences are inverted relative to the other? Further, does the robot express a preference for future pleasures over past ones, and past pains over future ones, as we do? Does it perhaps wonder whether *we humans* are conscious, despite our being constituted by different material from it? Although such interrogation can generate false negatives—just as conscious nonhuman animals and human infants would flunk the test—the likelihood of false positives could be reduced by "boxing in" the robots' knowledge base (Schneider 2020). The strength of ACT, however, is true positives: robots that pass seem very likely to be conscious.[5]

Suppose we have solid grounds for believing a robot is conscious. Is it also sentient? Since sentience involves the capacity for consciousness featuring pleasant or unpleasant experiences—in short, feelings—we could look for behaviors that seem to evince feelings. We might continue the ACT interview by asking the robot about its feelings in circumstances that seem likely to generate fear, anger, frustration, pleasure, pride, or the like. We might also ask it to describe, for example, the experienced difference between responding to an insult with a particular behavior and responding to the insult with the same behavior plus anger. What does the feeling add to the experience? If a robot affords us good reason to believe it is conscious, it would seem fairly easy to determine whether its consciousness includes pleasant or unpleasant experiences.

Suppose, however, we have good reason to believe a particular robot is conscious but *not* sentient. The machine apparently thinks, consciously, and has aims of a sort but cannot desire anything (in a sense of "desire" that implies caring about the object of desire), cannot experience any sensory inputs as pleasant or unpleasant, and cannot have any moods or emotional states. Because this strange, hyperbolically stoical entity lacks interests, nothing can be done for its sake. It lacks moral status.

[5] A third approach deploys the integrated information theory (IIT) of consciousness, developed by Giulio Tononi and colleagues (2016), as a basis for determining whether a particular being or entity is conscious. Despite some undeniable strengths, IIT is a controversial theory of the nature of consciousness, and other things being equal, it is preferable if tests for consciousness do not rest on controversial theories. For a general overview of the epistemological issues, see Shevlin 2021.

Some will deny my inference that insentient beings would necessarily lack interests and therefore moral status. They might ask us to imagine angel-like beings who, although entirely incapable of pleasant or unpleasant experiences, nevertheless had a sort of preference to act in accordance with the moral law or in accordance with their moral (or other) values. If one interfered with their attempts to act in this way, one would thwart their interest in doing so. This, according to the argument, suggests that even an insentient being could have interests if that being were conscious and had values.[6]

This very interesting challenge might have implications for any robots that achieve consciousness but not sentience. However, I continue to hold, tentatively, that sentience is required for interests. In my view, the angel-like creatures just described either *do* care about the completion of their aims and accordingly tend to feel some (unpleasant) frustration at their thwarting, in which case they are sentient and have interests; or, if they really lack all such feelings, then their complete emotional indifference and lack of feeling more generally make them entirely invulnerable such that they lack interests. The latter point is consonant with an attractive hypothesis: that much of the point of ascribing moral status is to note the *vulnerability* of certain beings as well as the moral importance of being responsive to their vulnerability. If one is totally invulnerable, on the present view, then one lacks moral status: a perfect, omnipotent God, although worthy of reverence, would not possess moral status (just as one who lacks a body has no need for clothing, food, or shelter).

Any readers who are unpersuaded by my reply to the present challenge may make a friendly amendment to my view, as follows: *in order to have interests and moral status a being must, first, be capable of consciousness; in addition, the being must either be sentient or (like the imagined angel-like beings) have values.*

Sentient Robots with Rights?

Suppose we believed certain robots were sentient. Then we should grant them moral status, which would entitle them to considerate treatment. For example, there would be a significant presumption against causing them to suffer, just as there is a significant moral presumption against causing sentient animals to suffer in laboratories or other settings. Further, if these robots lived with us and were dependent on us much as companion animals are, we ought to look after their basic needs, meaning the basic conditions underlying their ability to maintain their physical integrity, functioning, and a decent experiential welfare. From an ethical perspective, the importance of looking after dependent robots' basic needs would seem even more important if they were laboring for our benefit.

Would sentient robots have rights? For purposes of discussion, let's assume that sentient individuals who have narrative identities—whom we classify as *per-*

[6] Frances Kamm pressed roughly this challenge on me in personal communications.

sons—have rights that generally protect against sacrificing their interests for the common good. Then we would need to ask, of particular robots, whether they have a sufficiently rich self-awareness to think of their existence as constituting a sort of story with distinguishable parts. What sorts of evidence might convince us? Such behaviors might include spontaneous self-referential speech, nuanced statements about the robot's own future or past, or questions about the purpose of its existence and place in the universe. If any robot persuades us that it has a narrative identity, we should grant it not only (nontrivial) moral status but also the stronger protection of rights. That would mean, for example, as it does with humans, that an expectation of maximizing utility is insufficient to justify ending the existence of a robot or harming it in any other fundamental way. (For a thoughtful discussion of what rights robots might have, see Liao 2020.)

This same practical conclusion would follow from a different account of the basis of rights, so long as rights conferred special protections on top of basic moral status and the evidence persuaded us that robots made the relevant grade. Critical here is not my particular understanding of personhood in relation to narrative identity, or even of personhood as grounding rights. What's most important is that we be prepared to deploy a reasonable conception of the basis of rights when the internal complexity, behavior, and achievements of robots make the question of their moral status a live issue.

Autonomous Robots?

Might robots be not only sentient beings, and persons with narrative identities, but also autonomous agents? Many human beings, such as elementary school children, are clearly persons yet lack the capacity for autonomous decision-making. For this reason, paternalistic protection of them is often appropriate where it would be inappropriate in the case of competent adults. Like such humans, robots with narrative identities should be ascribed moral rights, but if they are incapable of substantially autonomous decision-making, their rights would not include those that deflect benign paternalism. So human caretakers could boss robots around for their own good. They could also enroll them as participants in minimal-risk, nontherapeutic research, just as human parents may do for their children.

According to the conception of autonomy I favor, an agent can act autonomously if and only if she can act (1) intentionally, (2) with sufficient understanding, (3) sufficiently freely of controlling influences, and (4) in light of her own values (DeGrazia and Millum 2021). My speculation is that a robot that affords us good reason to believe it is conscious, sentient, and narratively self-aware might very well meet these conditions.

One might deny that a robot could act sufficiently freely of controlling influences, as condition (3) requires. After all, mustn't a robot follow its human-created program? Let's set aside the intriguing possibility that advanced robots will write the programs for future generations of robots. Even among human-designed robots, some already learn through their experience—using a self-replicating neural network—and deploy their growing knowledge to solve problems or perform tricky tasks. (See, for example, DigInfo 2011). With major advances in robot learning, many robots will constantly modify their own programs, just as each human being modifies, through learning, the "program" she was afforded through genetics and environmental factors. I assume that competent adult human beings can act autonomously and, by parity of reasoning, this seems entirely possible for robot persons. The key insight is that robots that learn through their experience and make decisions on the basis of their evolving knowledge store thereby transcend, in an important sense, their human-made programs.

Can individual robots have values, as required in the above analysis of autonomy? Yes. A future robot might value, say, its own survival, performing its jobs well, and protection of humanity. Drawing from Isaac Asimov's classic novel *I, Robot* (1950), which features three value-employing laws of robotics to which all robots are bound, one might object that if robots have values, it's only because they were programmed to have them. But that does not seem necessarily true, any more than you and I can only have the values we were "programmed" to have by our biology and environmental influences; moreover, the issue is whether robots can *have* values, not how they acquire them. Once we acknowledge that robots can learn from experience and make their own decisions intelligently and flexibly—and there seem to be no good grounds for denying this possibility—we should judge that sufficiently advanced robots may be as capable of autonomy as we are (see Peterson 2017). If so, we will have to treat autonomous robots as free agents. They will have a right to refuse to participate in research, a right to decline the robot analog of medical treatment, and many other autonomy rights.

It takes little reflection to realize that treating certain robots as free agents could radically change human society. Indeed, the term "*human* society" might prove inadequate, considering what society would become if robots were recognized as having autonomy rights and possibly eligible for citizenship, with full protection of the law. Rather than explore the rich details of this thought-experiment, my intention here is to note how momentous such a development could be. Those who regard these possible developments as socially destabilizing or overly dangerous might prefer that humanity be careful never to create robots that might become autonomous—if it is possible to prevent such developments.

The New Speciesism and a Practical Dilemma for Robotics

Advances in robotics and AI are stimulating a new area of applied ethics—sometimes called "roboethics." One finds lectures, op-eds, and journal articles on such issues as: what criteria should autonomous vehicles apply in trolley-problem-like situations; who bears responsibilities when such vehicles cause fatalities; in accordance with what moral principles should military robots be deployed; and so on (see, for example, Jones, Kaufman, and Edenberg 2018; Nyholm 2018; Nyholm and Smids 2016). This article has focused on a single cluster of issues, considered before the time when they will seem practically relevant: what properties must robots have in order to have moral status, rights, and autonomy rights in particular? And how can we know, or justifiably believe, they possess the relevant properties?

Reflections about the possible moral status of future robots suggest that we are approaching a new struggle involving speciesism. The familiar struggle concerns nonhuman animals. Speciesists (as I use the term) maintain that membership in *Homo sapiens* per se confers unique or radically superior moral status. They also tend to support, in practice if not also in principle, the routine exploitation of animals for human purposes, including relatively trivial purposes such as convenience, marginal gains in dietary pleasure, and entertainment. Anti-speciesists, including most scholars working in animal ethics, deny that species per se bears on moral status and turn a critical eye on many mainstream human uses of animals.

The new battleground will feature a new type of speciesist: one who denies that artificial entities, regardless of their capacities or other properties, can have moral status or rights. Some might argue that, because robots are not alive, they will have less moral status than living things who are otherwise relevantly similar to robots. While such a thinker would avoid speciesism, she would nevertheless be guilty of what we might call *biologism*—an unwarranted belief that, other things being equal, life per se confers higher moral status. One might not find this belief so objectionable, considering that we deny human corpses have the moral status of living human beings. But this fact is better explained by the observation that corpses are entirely devoid of mental life, including sentience, and therefore lack interests. If you could convince me that a dead body were sentient, you would convince me that it had moral status. What is relevant here is not life but the possession of interests.

The new speciesism might manifest itself not only in irrational denials that sentient robots would have moral status or that robot persons would have rights, but also in claims about what robots can and cannot achieve. For example, some people will claim that, because robots are made by humans, they are incapable of creativity (du Sautoy 2019). This is mistaken. It is like saying that, if it turns out that we human beings are created by God, then we cannot be creative. But we often are creative. If one denies that robots can be creative because they have only *derived* intentionality (goals, desires), unlike the natural intentionality of their human creators (Searle 1984), I reply that robots that learn from experience and make their own decisions partly in light of their experience will be just as capable of forming their own specific intentions—including what to create—as we are.

Advances in robotics are driven both by scientific inquisitiveness and by corporate interests. The latter may drive most of the financial investments needed for major progress in the field. The possibility that such progress will usher in the day when robots have moral status presents a dilemma to corporations and to humanity at large. We will want robots to perform certain tasks (such as medical diagnoses or weather predictions) better and faster than humans can perform them, to remove us from certain dangers (as with bomb deactivation), or just to spare us from certain burdens (say, cleaning the house and caring for Grandpa).

But, if robots doing this work are sentient, they will have their own feelings and interests, and their interests would merit serious moral consideration. And if these robots have narrative identities, they should be ascribed rights that will ordinarily block appeals to utility as justifications for using them in ways contrary to their interests. Further, if they can act autonomously, they will have a right to pursue their own life-plans—or rather, "existence-plans." In such a case, the very progress that might deliver the practical advantages we seek from robotics might accidentally motivate legitimate claims that we are wrongfully exploiting, or even enslaving, the artificial beings we have created.[7] No one should be surprised if the new field of roboethics, developing in parallel with robotics itself, leads to manifestos with such titles as *Robot Liberation* or *The Case for Robot Rights*.[8]

References

Agar, N. 2020. "How to Treat Machines that Might Have Minds." *Philos Technol* 33: 269–82.

Asimov, I. 1950. *I, Robot*. New York: Gnome Press.

Armstrong, S., and R. Botzler. 2017. *The Animal Ethics Reader*, 3rd ed. New York: Routledge.

Basl, J. 2014. "Machines as Moral Patients We Shouldn't Care About [Yet]: The Interests and Welfare of Current Machines." *Philos Technol* 27: 79–96.

Basl, J., and J. Bowen. 2020. "AI as a Moral Right-Holder." In *The Oxford Handbook of Ethics of AI*, ed. M. Dubber, F. Pasquale, and S. Das, 277–94. New York: Oxford University Press.

Beauchamp, T. L., and R. G. Frey. 2011. *The Oxford Handbook of Animal Ethics*. New York: Oxford University Press.

Block, N. 1995. "On a Confusion about the Function of Consciousness." *Behav Brain Sci* 18: 227–24.

[7] In "Robots Should Be Slaves" (2010), Joanna Bryson argues that we should be careful *not* to create sentient robots—in which case the robot "slaves," on my account, would simply be complicated tools, not beings with moral status. Steve Peterson (2017) entertains the possibility that we could create robot persons to be our dedicated servants without wronging them by designing them to find fulfillment in serving our needs, allowing them to have leisure time in which to pursue other aims, and the like.

[8] Compare the titles of Singer's *Animal Liberation* (1975) and Regan's *The Case for Animal Rights* (1983).

Bostrom, N., and E. Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, ed. W. Ramsey and K. Frankish, 316–32. Cambridge: Cambridge University Press.

Bryson, J. 2010. "Robots Should be Slaves." In *Close Engagements with Artificial Companions*, ed. Y. Wilks, 63–74. Amsterdam: John Benjamins.

Clarke, A. C. 1968. *2001: A Space Odyssey*. New York: New American Library.

Coeckelbergh, M. 2010. "Robot Rights? Towards a Social-Relational Justification for Moral Consideration." *Ethics Info Technol* 12: 209–21.

DeGrazia, D. 1996. *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.

DeGrazia, D. 2008. "Moral Status as a Matter of Degree?" *South J Philos* 46: 181–98.

DeGrazia, D., and J. Millum. 2021. *A Theory of Bioethics*. Cambridge: Cambridge University Press.

DigInfo. 2011. "Robert Learns, Thinks, and Acts by Itself." DigInfo TV, Aug. 1. https://youtu.be/OC2TTslf_YM.

du Sautoy, M. 2019. "Can AI Ever Be Truly Creative?" *New Scientist* (11 May): 38–41.

Gellers, J. 2021. *Rights for Robots*. Oxford: Routledge.

Gunkel, D. 2018. "The Other Question: Can and Should Robots Have Rights?" *Ethics Info Technol* 20: 87–99.

Jaworska, A., and J. Tannenbaum. 2018. "The Grounds of Moral Status." In *Stanford Encyclopedia of Bioethics*, ed. E. Zalta. https://plato.stanford.edu/entries/grounds-moral-status/.

Jones, M. L., E. Kaufman, and E. Edenberg. 2018. "AI and the Ethics of Automating Consent." *IEEE* 16 (3): 64–72.

Kittay, E. F. 2005. "At the Margins of Moral Personhood." *Ethics* 116: 100–131.

Kripke, S. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

Levy, D. 2009. "The Ethical Treatment of Artificially Conscious Robots." *Int J Soc Robotics* 1: 209–16.

Liao, S. M. 2020. "The Moral Status and Rights of Artificial Intelligence." In *Ethics of Artificial Intelligence*, ed. S. M. Liao, 480–503. New York: Oxford University Press.

Nagel, T. 1974. "What is it Like to be a Bat?" *Philos Rev* 83: 435–50.

Nyholm, S. 2018. "Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci." *Sci Eng Ethics* 24: 1201. DOI: 10.1007/x11948-017-9943-x.

Nyholm, S., and J. Smids. 2016. "The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?" *Ethic Theor Moral Pract* 19: 1275–89.

Peterson, S. 2017. "Designing People to Serve." In *Robot Ethics*, ed. P. Lin, K. Abney, and G. Bekey, 283–98. New York: Oxford University Press.

Regan, T. 1983. *The Case for Animal Rights*. Berkeley: University of California Press.

Schneider, S. 2020. "How to Catch an AI Zombie: Testing for Consciousness in Machines." In *Ethics of Artificial Intelligence*, ed. S. M. Liao, 439–58. New York: Oxford University Press.

Schwitzgebel, E., and M. Garza. 2015. "A Defense of the Rights of Artificial Intelligences." *Midwest Stud Philos* 39: 89–119.

Searle, J. 1984. *Minds, Brains, and Science*. Cambridge: Harvard University Press.

Shevlin, H. 2021. “How Could We Know When a Robot Was a Moral Patient?” *Cambridge Q Healthc Ethics* 30: 450–71.

Singer, P. 1975. *Animal Liberation*. New York: Avon.

Steinbock, B. 2011. *Life Before Birth*, 2nd ed. New York: Oxford University Press.

Taylor, P. 1986. *Respect for Nature*. Princeton: Princeton University Press.

Tononi, G., et al. 2016. “Integrated Information Theory: From Consciousness to its Physical Substrate.” *Nat Rev Neurosci* 17: 450–61.

Varner, G. 1998. *In Nature’s Interests?* New York: Oxford University Press.

Warren, M. A. 1997. *Moral Status*. Oxford: Oxford University Press.