Author's pre-print. *The Oxford Handbook of Ethics of AI* (pp. 539-553), ed. by Markus D. Dubber, Frank Pasquale and Sunit Das. New York: Oxford University Press, 2020. ISBN: 978-0190067397

CHAPTER 28

PERSPECTIVES ON ETHICS OF AI

Philosophy

David J. Gunkel – Northern Illinois University (USA)

Whether we recognize it as such or not, we are in the midst of an AI invasion. The machines are everywhere and doing virtually everything. As these various devices and systems come to occupy influential positions in contemporary culture—positions where they are not necessarily mere tools or instruments of human action but a kind of social entity in their own right—we will need to ask ourselves some rather interesting but difficult questions: At what point might an AI, an algorithm, or other autonomous system be held accountable for the decisions it makes or the actions it initiates? When, if ever, would it make sense to say, "It's the computer's fault"? Conversely, when might an intelligent artifact or other socially interactive mechanism be due some level of social standing or respect? When, in other words, would it no longer be considered nonsense to inquire about the standing of artifacts and to ask the question: "Can and should AI have rights?"

My own response to these questions takes the form of a question, something that I have called *The Machine Question.* And this mode of response has, as one might anticipate, received some criticism for answering a question with a question.[1] I prefer, however, to read this criticism positively, and I do so because *questioning* is the defining condition of the philosophical endeavor. Philosophers as varied as Martin Heidegger, Daniel Dennett, George Edward Moore, and Slavoj Žižek have all, at one time or another, argued that the principal objective of philosophy is not to supply answers to difficult questions but to examine the questions themselves and our modes of inquiry. "The task of philosophy," Žižek writes, "is not to provide answers or solutions, but to submit to critical analysis the questions themselves, to make us see how the very way we perceive a problem is an obstacle to its solution."[2]

Following this procedure, this chapter demonstrates how and why the way we have typically perceived the problem of AI ethics is in fact a problem and an obstacle to its own solution. Toward this end, I will first demonstrate how the usual way of proceeding already involves considerable philosophical problems, and that these difficulties do not proceed from the complex nature of the subject matter that is asked about but derive from the very mode of inquiry. In other words, I will demonstrate how asking seemingly correct and intuitive questions might already be a significant problem and an obstacle to their solution. Second, and in response to this, I will advocate for an alternative mode of inquiry—another way of asking the question that is capable of accommodating the full philosophical impact and significance of AI. Third, the objective of the effort will be to respond to the question concerning AI not just as an opportunity to investigate the moral and social status of technological artifacts but as a challenge to rethink the basic configurations of moral philosophy itself.

1 Standard Operating Presumptions or The Default Setting

From a traditional philosophical perspective, the question concerning both the rights and responsibilities of AI would not only be answered in the negative but the query itself risks incoherence. As J. Storrs Hall has explained, "Morality rests on human shoulders, and if machines changed the ease with which things were done, they did not change the responsibilities for doing them. People have always been the only 'moral agents.' Similarly, people are largely the objects of responsibility. There is a developing debate over our responsibilities to other living creatures, or species of them....We have never, however, considered ourselves to have 'moral' duties to our machines, or them to us."[3] This statement sounds correct. Human beings design, develop, and deploy technology. For this reason, it is the human designer, manufacturer, or user who is responsible for the technology and what is eventually done (or not done) with it. Additionally, the only rights that would need to be respected in the process of using or applying a technology are those privileges, claims, powers, and/or immunities belonging to the other human persons who are on the receiving end and affected by the employment of a particular technological system or device.

This explanation is persuasive precisely because it is structured and informed by the answer that is typically provided for the question concerning technology. "We ask the question concerning technology," Martin Heidegger writes, "when we ask what it is. Everyone knows the

two statements that answer our question. One says: Technology is a means to an end. The other says: Technology is a human activity. The two definitions of technology belong together. For to posit ends and procure and utilize the means to them is a human activity."[4] According to Heidegger's analysis, the presumed role and function of any kind of technology—whether it be a simple hand tool, jet airliner, or a sophisticated robot—is that it is a means employed by human users for specific ends. Heidegger calls this particular characterization of technology "the instrumental definition" and indicates that it forms what is considered to be the "correct" understanding of any kind of technological contrivance.

The instrumental theory, therefore, "offers the most widely accepted view of technology. It is based on the common sense idea that technologies are 'tools' standing ready to serve the purposes of users."[5] And because an instrument or tool "is deemed 'neutral,' without valuative content of its own" a technological artifact is evaluated not in and of itself, but on the basis of the particular employments that have been decided by its human designer or user. "Computer systems," Deborah Johnson writes, "are produced, distributed, and used by people engaged in social practices and meaningful pursuits. This is as true of current computer systems as it will be of future computer systems. No matter how independently, automatic, and interactive computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision."[6] On this account, then, the bar for extending moral consideration to a machine, like an "intelligent" robot or AI, appears to be impossibly high if not insurmountable. In order for a technological artifact to have anything like independent moral status, it would need to be recognized as another subject and not just an object or instrument of human endeavor.

Standard approaches to deciding questions of moral subjectivity focus on what Mark Coeckelbergh calls "(intrinsic) properties." This method is rather straight forward and intuitive: identify one or more morally relevant properties and then find out if the entity in question has them or would be capable of having them.[7] In this transaction, ontology proceeds ethics; what something is determines how it is treated. Or as Luciano Floridi describes it "what the entity is determines the degree of moral value it enjoys, if any."[8] According to this standard procedure, the question concerning machine moral status would need to be decided by first identifying which property or properties would be necessary and sufficient for moral standing and then figuring out whether a particular AI or a class of AI possesses these properties or not. Deciding

things in this fashion, although entirely reasonable and expedient, has at least four critical difficulties.

1.1 Substantive Problems

How does one ascertain which exact property or properties are necessary and sufficient for moral status? In other words, which one, or ones, count? The history of moral philosophy can, in fact, be read as something of an on-going debate and struggle over this matter with different properties vying for attention at different times. And in this process many properties—that at one time seemed both necessary and sufficient—have turned out to be either spurious, prejudicial or both. Take for example a rather brutal action recalled by Aldo Leopold at the beginning of his essay "The Land Ethic": "When god-like Odysseus, returned from the wars in Troy, he hanged all on one rope a dozen slave-girls of his household whom he suspected of misbehavior during his absence. This hanging involved no question of propriety. The girls were property. The disposal of property was then, as now, a matter of expediency, not of right and wrong."[9] At the time Odysseus is reported to have done this, only male heads of the household were considered legitimate moral and legal subjects. Everything else—his women, his children, and his animals—were property that could be disposed of without any moral consideration whatsoever. But from where we stand now, the property "male head of the household" is clearly a spurious and rather prejudicial criteria for determining moral status.

Similar problems are encounter with, for example, rationality, which is the property that eventually replaces the seemingly spurious "male head of the household." When Immanuel Kant defined morality as involving the rational determination of the will, non-human animals, which do not (at least since the Cartesian *bête-machine*) possess reason, are immediately and categorically excluded from moral consideration. The practical employment of reason does not concern animals and, when Kant does make mention of animality, he only uses it as a foil by which to define the limits of humanity proper.[10] It is because the human being possesses reason, that he (and "human being," in this case and point in time, was principally defined as male) is raised above the instinctual behavior of a mere brute and able to act according to the principles of pure practical reason.

The property of reason, however, is contested by efforts in animal rights philosophy, which begins, according to Peter Singer, with a critical response issued by Jeremy Bentham:

"The question is not, 'Can they reason?' nor, 'Can they talk?' but 'Can they suffer?'"[11] For Singer, the morally relevant property is not speech or reason, which he believes sets the bar for moral inclusion too high, but sentience and the capability to suffer. In *Animal Liberation* and subsequent writings, Singer argues that any sentient entity, and thus any being that can suffer, has an interest in not suffering and therefore deserves to have that interest taken into account. Tom Regan, however, disputes this determination, and focuses his "animal rights" thinking on an entirely different property. According to Regan, the morally significant property is not rationality or sentience but what he calls "subject-of-a-life." Following this determination, Regan argues that many animals, but not all animals (and this qualification is important, because the vast majority of animal are excluded from his brand of "animal rights" thinking), are "subjects-of-a-life": they have wants, preferences, beliefs, feelings, etc. and their welfare matters to them. Although these two formulations of animal rights effectively challenge the anthropocentric tradition in moral philosophy, there remains considerable disagreement about which exact property is the necessary and sufficient condition for moral consideration.

1.2 Terminological Problems

Irrespective of which property (or set of properties) comes to be operationalized as the condition for moral standing, they each have terminological troubles insofar as things like rationality, consciousness, sentience, etc. mean different things to different people and seem to resist univocal definition. Consciousness, for example, is one of the properties that is often cited as a sufficient conditions for moral subjectivity. But consciousness is persistently difficult to define or characterize. The problem, as Max Velmans points out, is that this term unfortunately "means many different things to many different people, and no universally agreed core meaning exists."[12] In fact, if there is any general agreement among philosophers, psychologists, cognitive scientists, neurobiologists, AI researchers, and robotics engineers regarding consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept. As Rodney Brooks admits, "we have no real operational definition of consciousness," and for that reason, "we are completely prescientific at this point about what consciousness is."[13]

To make matters worse, the problem is not just with the lack of a basic definition; the problem may itself already be a problem. "Not only is there no consensus on what the term consciousness denotes," Güven Güzeldere writes, "but neither is it immediately clear if there

actually is a single, well-defined 'the problem of consciousness' within disciplinary (let alone across disciplinary) boundaries. Perhaps the trouble lies not so much in the ill definition of the question, but in the fact that what passes under the term consciousness as an all too familiar, single, unified notion may be a tangled amalgam of several different concepts, each inflicted with its own separate problems."[14] Although consciousness, as Anne Foerst remarks, is the secular and supposedly more "scientific" replacement for the occultish "soul," it turns out to be just as much an occult property or black box.[15]

Other properties do not do much better. Suffering and the experience of pain is just as nebulous, as Daniel Dennett cleverly demonstrates in the essay, "Why You Can't Make a Computer that Feels Pain." In this provocatively titled essay, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot."[16] At the end of what turns out to be a rather protracted and detailed consideration of the problem, Dennett concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect. According to Dennett, the reason you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is a product of the fact that we remain unable to decide what pain is in the first place. What Dennett demonstrates, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or a robot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and indeterminate. "There can," Dennett writes at the end of the essay, "be no true theory of pain, and so no computer or robot could instantiate the true theory of pain, which it would have to do to feel real pain."[17]

1.3 *Epistemological Problems*

As if responding to Dennett's challenge, engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses but also systems capable of evincing something that appears to be what we generally recognize as "pain."[18] The interesting problem in these cases is determining whether this is in fact "real pain" or just a simulation of pain. In other words, once the morally significant property or properties have been identified, how can one be entirely certain that a particular entity possesses it, and actually possesses it instead of merely

simulating it? Resolving this problem is tricky business, especially because most of the properties that are considered morally relevant tend to be internal mental or subjective states that are not immediately accessible or directly observable. As Paul Churchland famously asked: "How does one determine whether something other than oneself—an alien creature, a sophisticated robot, a socially active computer, or even another human—is really a thinking, feeling, conscious being; rather than, for example, an unconscious automaton whose behavior arises from something other than genuine mental states?"[19]

Though "pain" is not the direct object of his analysis, the epistemological difficulty of distinguishing between the "real thing" and its mere simulation is something that was addressed and illustrated by John Searle's "Chinese Room" thought experiment. "Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese."[20] The point of Searle's imaginative (albeit ethnocentric illustration) is quite simple—simulation is not the real thing. Merely shifting symbols around in a way that looks like linguistic understanding is not really an understanding of the language. A similar point has been made in the consideration of other properties, like sentience and the experience of pain. Even if, as J. Kevin O'Regan writes, it were possible to design an artifact that "screams and shows avoidance behavior, imitating in all respects what a human would do when in pain …. All this would not guarantee that to the robot, there was actually something it was like to have the pain. The robot might simply be going through the motions of manifesting its pain: perhaps it actually feels nothing at all."[21] The problem exhibited by both examples, however, is not simply that there is a difference between simulation and the real thing. The problem is that we remain persistently unable to distinguish the one from the other in any way that would be considered entirely satisfactory. "There is," as Dennett concludes, "no proving that something that seems to have an inner life does in fact have one—if by 'proving' we understand, as we often do, the evincing of evidence that can be seen to establish by principles already agreed upon that something is the case."[22]

1.4 Moral Problems

Finally the properties approach, when applied to humanly designed artifacts like AI, runs into ethical problems. Here is how Wendell Wallach and Colin Allen formulate it: "If (ro)bots might one day be capable of experiencing pain and other affective states a question that arises is whether it will be moral to build such systems—not because of how they might harm humans, but because of the pain these artificial systems will themselves experience. In other words, can the building of a (ro)bot with a somatic architecture capable of feeling intense pain be morally justified …?"[23] If it were in fact possible to construct a mechanism that is sentient and "feels pain" (however that term would be defined and instantiated in the device) in order to demonstrate the underlying ontological properties of the artifact, then doing so might be ethically suspect insofar as in constructing such a device we do not do everything in our power to minimize its suffering. For this reason, moral philosophers and AI scientists/engineers find themselves in a curious and not entirely comfortable situation. One would need to be able to construct an artifact that feels pain in order to demonstrate the actual presence of sentience; but doing so could be, on that account, already to risk engaging in actions that are immoral and that violate the rights of others.

The legal aspects of this problem are something that is taken up and addressed by Lantz Fleming Miller, who points out that efforts to build what he calls "maximally humanlike automata" (MHA) could run into difficulties with informed consent: "The quandary posed by such an MHA in terms of informed consent is that it just may qualify, if not precisely for a human being, then for a being meriting all the rights that human beings enjoy. This quandary arises from the paradox of its construction vis-à-vis informed consent: it cannot give its consent for the relevant research and development performed to ensure its existence."[24] According to Miller's argument, the very effort to construct a hypothetical MHA—an artifact that if not precisely human is at least capable of qualifying for many of the responsibilities and rights that human beings currently enjoy—already violates that entity's right to informed consent insofar as the mechanism would not have been informed about and given the opportunity to consent to its being constructed. There is, in other words, something of a moral paradox in trying to demonstrate machine moral standing, either now or in the future. In order to run the necessary demonstration and construct a system or device that could qualify for meriting human-level

moral respect, one would need to build something that not only cannot give consent in advance of its own construction but which also could retroactively (after having been created) withdraw consent to its having been fabricated in the first place. Consequently, there is a moral and/or legal problem involved in conducting this research: The demonstration that an AI is a legitimate moral subject with rights that would need to be duly respected might already violate the very rights that come to be demonstrated.

2 Thinking Otherwise, or The Relational Turn

In response to these problems, philosophers—especially in the continental tradition—have advanced alternative approaches that can be called "thinking otherwise."[25] This phrase signifies different ways to formulate the question concerning moral standing that is open to and able to accommodate others—and other forms of morally significant otherness. Contrary to the usual way of deciding things, these efforts do not endeavor to determine ontological criteria for inclusion or exclusion but begin from the existential fact that we always and already find ourselves in situations facing and needing to respond to others—not just other human beings but non-human animals, the environment, organizations, and technological artifacts, like AI. In fact, recent debates concerning the social status of corporations turn on the question whether moral and legal standing derive from intrinsic properties at all or are, as Anne Foerst, describes it, a socially constructed and conferred honorarium.[26]

What is important here, is that these alternatives shift the focus of the question and change the terms of the debate. Here it is no longer a matter of, for example, "Can AI be a moral subject?" which is largely an ontological query concerned with the prior discovery of intrinsic and morally relevant properties. Instead it is something like "Should AI be a moral subject?" which is an ethical question and one that is decided not on the basis of what things are but on how we relate and respond to them in actual social situations and circumstances. In this case the actual practices of social beings in relationship with each other take precedence over the ontological properties of the individual entities or their material implementations. This change in perspective provides for a number of important innovations that affect not just AI ethics but moral philosophy itself.

2.1 Relationalism

Moral status is decided and conferred not on the basis of subjective or internal properties but according to objectively observable, extrinsic relationships. "Moral consideration," as Mark Coeckelbergh describes it, "is no longer seen as being 'intrinsic' to the entity: instead it is seen as something that is 'extrinsic': it is attributed to entities within social relations and within a social context."[27] As we encounter and interact with others, this other entity is first and foremost situated in relationship to us. Consequently, the question of moral status does not necessarily depend on what the other is in its essence but on how she/he/it/they (and pronouns matter in this context) stand in relationship to us and how we decide, in the face of the other (to use Levinasian terminology), to respond. In this formulation, "relations are prior to the things related"[28], instituting what Anne Gerdes (following Coeckelbergh) calls "a relational turn" in ethics.[29]

This shift in perspective, it is important to point out, is not just a theoretical proposal made by "armchair philosophy"; it has been experimentally confirmed in a number of practical investigations. The computer as social actor (CASA) studies undertaken by Byron Reeves and Clifford Nass, for example, demonstrated that human users will accord computers social standing similar to that of another human person and this occurs as a product of the extrinsic social interaction, irrespective of the intrinsic properties (actually known or not) of the entities in question. "Computers, in the way that they communicate, instruct, and take turns interacting, are close enough to human that they encourage social responses. The encouragement necessary for such a reaction need not be much. As long as there are some behaviors that suggest a social presence, people will respond accordingly…. Consequently, any medium that is close enough will get human treatment, even though people know it's foolish and even though they likely will deny it afterwards."[30] These results have been verified in subsequent studies with social robots[31], explosive ordinance disposal (EOD) robots[32], and even mundane objects like the Roomba robotic vacuum clearer[33]. As Scheutz reports: "While at first glance it would seem that the Roomba has no social dimension (neither in its design nor in its behavior) that could trigger people's social emotions, it turns out that humans, over time, develop a strong sense of gratitude toward the Roomba for cleaning their home. The mere fact that an autonomous machine keeps working for them day in and day out seems to evoke a sense of, if not urge for, reciprocation."[34]

2.2 Radically Empirical

This approach is phenomenological or (if you prefer) radically empirical in its epistemological commitments. Because moral consideration is dependent upon extrinsic social circumstances and not prior determinations of internal properties, the seemingly irreducible problem of other minds is not some fundamental epistemological limitation that must be addressed and resolved prior to moral decision making. Instead of being derailed by the epistemological problem of other minds, this approach to moral thinking immediately affirms and acknowledges this difficulty as the basic condition of possibility for ethics as such. Consequently, "the ethical relationship," Emmanuel Levinas writes, "is not grafted on to an antecedent relationship of cognition; it is a foundation and not a superstructure…It is then more cognitive than cognition itself, and all objectivity must participate in it."[35] It is for this reason that Levinasian philosophy focuses attention not on other minds, but on the face of the other.[36] Or as Richard Cohen succinctly explains in what could be an advertising slogan for Levinasian thought: "Not other 'minds,' mind you, but the 'face' of the other, and the faces of all others."[37]

This also means that the order of precedence in moral decision making can and perhaps should be reversed. Internal properties do not come first and then moral respect follows from this ontological fact. We have things backwards. Instead the morally significant properties—those ontological criteria that we assume ground moral respect—are what Žižek (2008, 209) terms "retroactively (presup)posited"[38] as the result of and as justification for decisions made in the face of social interactions with others. In other words, we project the morally relevant properties onto or into those others who we have already decided to treat as being socially significant—those others who are deemed to possess face, in Levinasian terminology. In social situations, then, we always and already decide between "who" counts as morally significant and "what" does not and then retroactively justify these actions by "finding" the properties that we believe motivated this decision making in the first place. Properties, therefore, are not the intrinsic *a prior* condition of possibility for moral standing. They are *a posteriori* products of extrinsic social interactions with and in the face of others.

This is not some theoretical formulation; it is practically the definition of machine intelligence. Although the phrase "artificial intelligence" is the product of an academic conference organized by John McCarthy at Dartmouth College in 1956, it is Alan Turing's 1950 paper and its "game of imitation," or what is now routinely called "the Turing Test," that defines

and characterizes the field. According to Turing's stipulations, if a computer is capable of successfully simulating a human being in communicative exchanges to such an extent that the interrogator in the game cannot tell whether he is interacting with a machine or another human person, then that machine would, Turing concludes, need to be considered "intelligent." Or in Žižek's terms, if the machine effectively passes for another human person in communicative interactions, the property of intelligence would be "retroactively (presup)posited" for that entity, and this is done irrespective of the actual internal states or operations of the interlocutor, which are, according to the stipulations of Turing's game, unknown and hidden from view.

2.3 Altruistic

Because ethics transpires in the relationship with others or the face of the other, extending the scope of moral standing can no longer be about the granting of rights to others. Instead, the other, first and foremost, questions my rights and challenges my being here. According to Levinas, "the strangeness of the Other, his [SIC] irreducibility to the I, to my thoughts and my possessions, is precisely accomplished as a calling into question of my spontaneity, as ethics."[39] This interrupts and even reverses the power relationship enjoyed by previous forms of ethics. Here it is not a privileged group of insiders who then decide to extend rights to others, which is the basic model of all forms of moral inclusion or what Peter Singer calls a "liberation movement."[40] Instead the other challenges and questions the rights and freedoms that I assume I already possess. The principal gesture, therefore, is not the conferring rights on others as a kind of benevolent gesture or even an act of compassion but deciding how to respond to the other, who always and already places my rights and assumed privilege in question. Such an ethics is *altruistic* in the strict sense of the word. It is "of or to others."

Finally this altruism is not just open to others but must remain permanently open and exposed to other others. "If ethics arises," as Matthew Calarco writes, "from an encounter with an Other who is fundamentally irreducible to and unanticipated by my egoistic and cognitive machinations," then identifying the "'who' of the Other" is something that cannot be decided once and for all or with any certitude.[41] This apparent inability or indecision is not necessarily a problem. In fact, it is a considerable advantage insofar as it opens the possibility of ethics to others and other forms of otherness. "If this is indeed the case," Calarco concludes, "that is, if it is the case that we do not know where the face begins and ends, where moral considerability

begins and ends, then we are obligated to proceed from the possibility that anything might take on a face. And we are further obligated to hold this possibility permanently open."[42]

3 Outcomes and Conclusions

We appear to be living in that future Norbert Wiener predicted over 50 years ago in *The Human Use of Human Beings*: "It is the thesis of this book," Wiener wrote, "that society can only be understood through a study of the messages and the communication facilities which belong to it; and that in the future development of these messages and communication facilities, messages between man and machines, between machines and man, and between machine and machine, are destined to play an ever increasing part."[43] As our world becomes increasingly populated by intelligent, socially interactive artifacts—devices that are not just instruments of human action but designed to be a kind of social actor in their own right—we will need to grapple with challenging questions concerning the status and moral standing of these machinic others—these other kind of others. Although this has been one of the perennial concerns in science fiction, it is now part and parcel of our social reality.

In formulating responses to these questions we can obviously deploy the standard properties approach. This method has considerable historical precedent behind it and constitutes what can be called the default setting for addressing questions concerning moral standing. And a good deal of the current work in moral machines, machine ethics, AI ethics, and the ethics of AI follow this procedure. But this approach, for all its advantages, also has considerable difficulties: 1) substantive problems with inconsistencies in the identification and selection of the qualifying properties for determining moral status, 2) terminological troubles with the definition of the morally significant property or properties, 3) epistemological difficulties with detecting and evaluating these properties in another, and 4) moral complications caused by the fact that the research necessary to demonstrate moral status runs the risk of violating the rights of others.

This does not mean, it is important to point out, that the properties approach is somehow wrong, misguided, or refuted on this account. It just means that the properties approach—despite its almost unquestioned acceptance as a kind of standard operating procedure—has limitations and that these limitations are becoming increasingly evident in the face of technological artifacts—in the face of others who are and remain otherwise. To put it in Žižek's terms, the properties approach, although appearing to be the right place to begin thinking about and

resolving the question of machine moral standing, may turn out to be the "wrong question" and even an obstacle to its solution.

As an alternative, I have proposed an approach to addressing AI ethics and the ethics of AI that is situated and oriented otherwise. This alternative circumvents many of the problems encountered in the properties approach by arranging for an ethics that is relational, radically empirical, and altruistic. This other way of thinking is informed by and follows from recent innovations in moral philosophy: 1) Levinasian thought, which puts ethics before ontology, making moral philosophy first philosophy, and 2) various forms of environmental ethics, like that developed by J. Baird Callicott, who argues that it is the social relationship that precedes and takes precedence over the things related. This does not mean, however, that this alternative is a panacea or some kind of moral theory of everything. It just arranges for other kinds of questions and modes of inquiry that are more attentive to the very real situation in which we currently find ourselves.

To put it in terms derived from Immanuel Kant's first critique—Instead of trying to answer the question of machine moral standing by continuing to pursue the properties approach, we should test whether we might not do better by changing the question and the terms of the debate. Consequently, my objective has not been to resolve the question of moral standing once and for all, but to ask about and evaluate the means by which we have situated and pursued this inquiry. This is not a dodge or a cop out. It is the one thing that philosophers and philosophy are good for. "I am a philosopher not a scientist," Daniel Dennett writes at the beginning of one of his books, "and we philosophers are better at questions than answers. I haven't begun by insulting myself and my discipline, in spite of first appearances. Finding better questions to ask, and breaking old habits and traditions of asking, is a very difficult part of the grand human project of understanding ourselves and our world."[44]

For this reason, the questions concerning AI and ethics are not just another set of problems to be accommodated to and resolved by existing moral theories or lists of ethical principles. It is instead in the face of increasingly social and interactive artifacts that moral theory and practice also comes to be submitted to a thorough reevaluation and critical questioning. AI ethics, therefore, is not just moral philosophy applied to the new opportunities and challenges of AI; it also calls for and requires a thorough reformulation of moral philosophy for and in the face of these other kinds of (artificial) others.

Notes

[1] Jeffrey D. Gottlieb, "Questions Left Unanswered," *Ethics & Behavior* 23 (2013): 163-166.

[2] Slavoj Žižek, "Philosophy, the 'Unknown Knowns,' and the Public Use of Reason," *Topoi* 25 (2006): 137.

[3] J. Storrs Hall, "Ethics for Machines," *KurzweilAI.net* (5 July 2001) http://www.kurzweilai.net/ethics-for-machines.

[4] Martin Heidegger, *The Question Concerning Technology and Other Essays*, trans. by William Lovitt (New York: Harper & Row, 1977), 4-5.

[5] Andrew Feenberg, *Critical Theory of Technology* (Oxford: Oxford University Press, 1991), 5.

[6] Deborah Johnson, "Computer Systems: Moral Entities but not Moral Agents," *Ethics and Information Technology* 8, no. 4 (2006): 197.

[7] Mark Coeckelbergh, *Growing Moral Relations: Critique of Moral Status Ascription* (New York: Palgrave MacMillan, 2012), 13-14.

[8] Luciano Floridi, *The Ethics of Information* (Oxford: Oxford University Press, 2013), 116.

[9] Aldo Leopold, *A Sand County Almanac* (Oxford: Oxford University Press, 1966), 237.

[10] Immanuel Kant, *Critique of Practical Reason*, trans. by Lewis White Beck (New York: Macmillan, 1985).

[11] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press, 2005), 283.

[12] Max Velmans, *Understanding Consciousness* (London, UK: Routledge, 2000), 5.

[13] Rodney Brooks, *Flesh and Machines: How Robots Will Change Us* (New York: Pantheon Books, 2002), 194.

[14] Güven Güzeldere, "The Many Faces of Consciousness: A Field Guide," in *The Nature of Consciousness: Philosophical Debates* (Cambridge, MA: MIT Press, 1997), 7.

[15] Gregory Benford and Elisabeth Malartre, *Beyond Human: Living with Robots and Cyborgs* (New York: Tom Doherty, 2007), 162.

[16] Daniel Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology* (Cambridge, MA: MIT Press, 1998), 191.

[17] Ibid., 228.

[18] See for example: J. Bates, "The Role of Emotion in Believable Agents," *Communications of the ACM* 37 (1994): 122–125. B. Blumberg, P. Todd and M. Maes, "No Bad Dogs: Ethological Lessons for Learning. In *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior* (Cambridge, MA: MIT Press, 1996): 295–304. Cynthia Breazeal and Rodney Brooks, "Robot Emotion: A Functional Perspective," in *Who Needs Emotions: The Brain Meets the Robot* (Oxford: Oxford University Press, 2004), 271–310.

[19] Paul Churchland, *Matter and Consciousness* (Cambridge, MA: MIT Press, 1999), 67.

[20] John Searle, "The Chinese Room, in *The MIT Encyclopedia of the Cognitive Sciences*, edited by R. A. Wilson and F. Keil (Cambridge, MA: MIT Press, 1999), 115.

[21] J. Kevin O'Regan "How to Build Consciousness into a Robot: The Sensorimotor Approach," in *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, edited by Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer (Berlin: Springer-Verlag, 2007), 332.

[22] Dennett, *Brainstorms*, 172.

[23] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: Oxford University Press, 2009), 209.

[24] Lantz Fleming Miller, "Responsible Research for the Construction of Maximally Humanlike Automata: The Paradox of Unattainable Informed Consent." *Ethics and Information Technology*. Published ahead of print (July 2017): 8.

[25] David J. Gunkel. *Thinking Otherwise: Philosophy, Communication, Technology* (West Lafayette, IN: Purdue University Press, 2007).

[26] Benford and Malartre, 165.

[27] Mark Coeckelbergh, "Robot Rights? Towards a Social-Relational Justification of Moral Consideration," *Ethics and Information Technology* 12 (2010): 214.

[28] J. Baird Callicott, *In Defense of the Land Ethic: Essays in Environmental Philosophy* (Albany, NY: SUNY Press, 1989), 110.

[29] Anne Gerdes, "The Issue of Moral Consideration in Robot Ethics." *ACM SIGCAS Computers & Society* 45 (2015): 274.

[30] Byron Reeves and Clifford Nass *The Media Equation* (Cambridge: Cambridge University Press, 1996), 22.

[31] Astrid Rosenthal-von der Pütten et al., "An Experimental Study on Emotional Reactions Towards a Robot," *International Journal of Social Robotics* 5 (2013): 17-34. Yutaka Suzuki et al., "Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography," *Scientific Reports* 5 (2015).

[32] Julie Carpenter, *Culture and Human-Robot Interaction in Militarized Spaces: A War Story* (New York: Ashgate, 2015).

[33] Ja-Young Sung, "My Roomba is Rambo: Intimate Home Appliances," in *Proceedings of UbiComp 2007* (Berlin: Springer-Verlag, 2007): 145-162.

[34] Matthias Scheutz, "The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics* (Cambridge, MA: MIT Press, 2012), 213.

[35] Emmanuel Levinas, *Collected Philosophical Papers*, trans. by Alphonso Lingis (Dordrecht: Martinus Nijhoff, 1987), 56.

[36] This particular use of Levinas's work require some qualification. Whatever the import of his unique contribution, Other in Levinas is still and unapologetically characterized as human. Although he is not the first to identify it, Jeffrey Nealon provides what is perhaps one of the most succinct descriptions of this problem in *Alterity Politics* (Durham, NC: Duke University Press, 1998): "In thematizing response solely in terms of the human face and voice, it would seem that Levinas leaves untouched the oldest and perhaps most sinister unexamined privilege of the same: *anthropos* [άνθρωπος] and only *anthropos*, has *logos* [λόγος]; and as such, *anthropos* responds not to the barbarous or the inanimate, but only to those who qualify for the privilege of 'humanity,' only those deemed to possess a face, only to those recognized to be living in the *logos*" (Nealon 1998, 71). If Levinasian philosophy is to provide a way to formulate an ethics that is able to respond to and to take responsibility for other forms of otherness we will need to use and interpret Levinas's own philosophical innovations in excess of and in opposition to him. Such efforts at "radicalizing Levinas," as Peter Atterton and Matthew Calarco (*Radicalizing Levinas*. Albany, NY: State University of New York Press, 2010) call it, take up and pursue Levinas's moral innovations in excess of the rather restricted formulations that he and his advocates and critics have typically provided. As Calarco in *Zoographies: The Question of the Animal from Heidegger to Derrida* (New York: Columbia

University Press, 2008, 55) explains, "Although Levinas himself is for the most part unabashedly and dogmatically anthropocentric, the underlying logic of his thought permits no such anthropocentrism. When read rigorously, the logic of Levinas's account of ethics does not allow for either of these two claims. In fact, as I shall argue, Levinas's ethical philosophy is, or at least should be, committed to a notion of universal ethical consideration, that is, an agnostic form of ethical consideration that has no a priori constraints or boundaries."

[37] Richard Cohen, *Ethics, Exegesis, and Philosophy: Interpretation After Levinas* (Cambridge: Cambridge University Press, 2001), 336.

[38] Slavoj Žižek, *For They Know Not What They Do: Enjoyment as a Political Factor* (London: Verso, 2008), 209.

[39] Emmanuel Levinas, *Totality and Infinity*, trans. by Alphonso Lingis (Pittsburgh, PA: Duquesne University, 1969), 43.

[40] Peter Singer, "All Animals are Equal," in *Animal Rights and Human Obligations* (New Jersey: Prentice-Hall, 1989), 148.

[41] Matthew Calarco, *Zoographies: The Question of the Animal from Heidegger to Derrida* (New York: Columbia University Press, 2008), 71.

[42] Ibid.

[43] Norbert Wiener, *The Human Use of Human Beings* (New York: Da Capo, 1954), 16.

[44] Daniel Dennett, *Kinds of Minds: Toward an Understanding of Consciousness* (New York: Basic Books, 1996), vii.

Bibliography

Anderson, Michael and Susan Leigh Anderson (eds). *Machine Ethics*. Cambridge: Cambridge University Press, 2011.

Asaro, Peter and Wendell Wallach (eds). *Machine Ethics and Robot Ethics*. New York: Routledge, 2016.

Coeckelbergh, Mark. *Growing Moral Relations: Critique of Moral Status Ascription*. New York: Palgrave MacMillan, 2012.

Dennett, Daniel C. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press, 1998.

Gunkel, David J. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* Cambridge, MA: MIT Press, 2012.

Lin, Patrick, Keith Abney and George A. Bekey (eds). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, 2012.

Reeves, Byron and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge: Cambridge University Press, 1996.

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

Turner, Jacob. *Robot Rules: Regulating Artificial Intelligence*. New York: Palgrave Macmillan, 2018.

Tzafestas, Spyros G. *Roboethics: A Navigating Overview*. New York: Springer, 2016.

Wallach, Wendell and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2009.

Wiener, Norbert. *The Human Use of Human Beings: Cybernetics and Society*, Boston, MA: Da Capo Press, 1988.