Chapter 3 from *Robo-Philosophy: Philosophy of, for, and by Social Robotics*.
Edited by Johanna Seibt, Raul Hakli and Marco Nørskov. MIT Press, 2016.

The Machine Question:

Rethinking Ethics in the Face of Others

Whether we acknowledge it or not, we are, in fact, in the midst of a robot invasion. The machines are everywhere and doing everything. They may have begun by displacing workers on the factory floor, but they now actively participate in many aspects of intellectual, social, and cultural life. This incursion is not some future possibility coming from a distant alien world. It is here. It is now. And resistance appears to be futile. As these increasingly social interactive devices come to occupy positions where they are not just tools or instruments of human conduct but a kind of social actor in their own right—we will need to ask ourselves important but rather difficult questions:

At what point might a robot—defined broadly and including not only physically embodied mechanisms like Baxter, Nao, and Pepper but also software implementations of artificial intelligence and deep learning like IBM's Watson and Google's TensorFlow—be held responsible for the decisions it makes or the actions it deploys? When, in other words, would it make sense to say "It's the computer's fault?" Likewise, at what point might we have to seriously consider extending something like rights—again defined broadly so as to include some level of civil, moral, or legal standing—to these socially aware and interactive devices? When, in other words, would it no longer be considered nonsense to suggest something like "the vindication of the rights of machines?"[1]

My own response to these questions takes the form of a question, something that I have called *The Machine Question* (Gunkel, 2012). And I have, as one might anticipate, received some criticism for responding to a question with a question (Allen, 2013; Gottlieb, 2013). I prefer, however, to read this criticism positively, and I do so because *questioning* is a specifically philosophical endeavor. Philosophers as varied as Martin Heidegger (1962), Daniel Dennett (1996), George Edward Moore (2005), and Slavoj Žižek (2006) 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 (2006, 137) 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."

Following this procedure, this chapter demonstrates how and why the way we have typically perceived the problem of ethics in the face of the machine 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—the usual method of asking about and addressing the question—already involves considerable philosophical problems, and that these difficulties proceed not from the complex nature of the subject matter that is asked about but 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 its 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 socially interactive machines. This alternative will capitalize on meta-ethical innovations coming out of continental philosophy, especially the work of Emmauel Levinas and others who follow his innovative lead. In the end, the objective of this effort is to respond to the question concerning social robotics not just as an opportunity to investigate the moral and social status of machines but as a challenge to rethink the basic configurations of moral philosophy itself. The *machine question*, therefore, is not only a matter of asking whether or not machines can or should have rights [2]; it is also a question about how we have formulated the concept of moral standing and determined who or what is included in the community of moral subjects.

1 Standard Operating Presumptions or The Default Setting

From a traditional philosophical perspective, the question of machine moral standing or robot rights not only would be answered in the negative but the query itself risks incoherence. “To many people,” David Levy (2005, 393) writes, “the notion of robots having rights is unthinkable.” It is unthinkable because robots [3], no matter how sophisticated or autonomous they might appear to be, are assumed to be nothing more than instruments of human activity and have no independent moral status whatsoever. As Barbara Johnson explains from the perspective of computer ethics: “Computer systems 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” (Johnson 2006, 197).

This rather intuitive and common sense view is structured and informed by the answer that is typically provided for the question concerning technology.

> We ask the question concerning technology 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. The manufacture and utilization of equipment, tools, and machines, the manufactured and used things themselves, and the needs and ends that they serve, all belong to what technology is (Heidegger 1977, 5).

According to Heidegger's analysis, the presumed role and function of any kind of technology, whether it be the product of handicraft or industrialized manufacture, is that it is a means employed by human users for specific ends. Heidegger terms 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.

Consequently, “the instrumentalist theory,” as Andrew Feenberg (1991, 5) summarizes it, “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." 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. On this account, the bar for extending moral consideration to a machine, like a socialable interactive robot, 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 a tool or an instrument of human endeavor.

Standard approaches to deciding these questions of moral subjectivity focus on what Mark Coeckelbergh (2012, 13) 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. In this transaction, ontology proceeds ethics. What something is determines how it is. Or as Luciano Floridi (2013, 116) describes it “what the entity is determines the degree of moral value it enjoys, if any…” According to this “'substance-attribute' ontology” (Seibt 2015, 4), the question concerning machine moral standing 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 robot or a class of robots 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

First, 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 (1966, 237) 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." 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—their women, their children, and their 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 standing.

Similar problems are encounter with, for example, the property of rationality, which is the property that eventually replace the seemingly spurious "male head of the household." When Immanuel Kant (1985, 17) defined morality as involving the rational determination of the will, non-human animals, which do not (at least since the Cartesian *bete-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 (*Tierheit*), he only uses it as a foil by which to define the limits of humanity proper. It is because the human being possesses reason, that he (and the human being, in this case, was principally male) is raised above the instinctual behavior of a mere brute and able to act according to the principles of pure practical reason (Kant 1985, 63).

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 (2005, 283): ''The question is not, 'Can they reason?' nor, 'Can they talk?' but 'Can they suffer?''' 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* (1975) 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" (Regan 1983, 243). 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"), are "subjects-of-a-life": they have wants, preferences, beliefs, feelings, etc. and their welfare matters to them (Regan 1983). 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

Second, irrespective of which property (or set of properties) is operationalized, 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 necessary conditions for moral agency (Himma 2009, 19). But consciousness is persistently difficult to define or characterize. The problem, as Max Velmans (2000, 5) points out, is that this term unfortunately "means many different things to many different people, and no universally agreed core meaning exists." 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 (2002, 194) admits, "we have no real operational definition of consciousness," and for that reason, "we are completely prescientific at this point about what consciousness is."

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 (1997, 7) 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." Although

consciousness, as Anne Foerst remarks, is the secular and supposedly more "scientific" replacement for the occultish "soul" (Benford and Malartre 2007, 162), it turns out to be just as much an occult property.

Other properties do not do much better. Suffering and the experience of pain—which is the property usually deployed in non-standard patient-oriented approaches like animal rights philosophy—is just as nebulous, as Daniel Dennett cleverly demonstrates in the essay, "Why You Cannot 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" (Dennett 1998, 191). 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 (1998, 228) 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." What Dennett proves, then, is not an inability to program a computer to "feel pain" but our initial and persistent inability to decide and adequately articulate what constitutes the experience of pain in the first place.

1.3 *Epistemological Problems*

As if responding to Dennett's challenge, robotics engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses (Bates 1994; Blumberg, Todd, and Maes 1996; Breazeal and Brooks 2004), like the dental-training robot Simroid "who" cries out in pain when students "hurt" it (Kokoro 2009), but also systems capable of evincing something that appears to be what we generally recognize as "pain." The interesting problem in these cases is determining whether this is in fact "real pain" or just a simulation of pain [4]. 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 "simulation 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 (1999, 67) 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?”

This is, of course, what philosophers call the other minds problem. Although this problem is not necessarily intractable, as I think Steve Torrance (2013) has persuasively argued, the fact of the matter is we cannot, as Donna Haraway (2008, 226) describes it, "climb into the heads of others to get the full story from the inside." And the supposed solutions to this "other minds problem," from reworkings and modifications of the Turing Test (Sparrow 2004) to functionalist approaches that endeavor to work around this problem altogether (Wallach and Allen 2009), only make things more complicated and indeterminate. "There is," as Dennett (1998, 172) points out, "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." Although philosophers, psychologists, and neuroscientists throw considerable argumentative and experimental effort at this problem, it is not able to be resolved in any way approaching what would pass for definitive evidence, strictly speaking. In the end, not only are these tests unable to demonstrate with any certitude whether animals, machines, or other entities are in fact conscious and therefore legitimate moral persons (or not), we are left doubting whether we can even say the same for other human beings. As Ray Kurzweil (2005, 380) candidly concludes, "we assume other humans are conscious, but even that is an assumption," because "we cannot resolve issues of consciousness entirely through objective measurement and analysis (science)."

1.4 Methodological Problems

Finally any decision concerning qualifying properties is necessarily a normative operation and an exercise of power over others. In making a determination about the criteria for moral inclusion, someone or some group universalizes their particular experience or situation and imposes this decision on others as the fundamental condition for moral consideration. “The institution of any practice of any criterion of moral considerability,” Thomas Birch (1993, 317) writes, “is an act of power over, and ultimately an act of violence toward, those others who turn out to fail the test of the criterion and are therefore not permitted to enjoy the membership benefits of the club of *consideranda.*” In other words, every criteria of moral inclusion, every “comprehensive list” of qualifying properties, no matter how neutral, objective, or universal it appears, is an imposition of power insofar as it consists in the

universalization of a particular value or set of values made by someone from a particular position of power. "The nub of the problem with granting or extending rights to others," Birch (1995, 39) concludes, "a problem which becomes pronounced when nature is the intended beneficiary, is that it presupposes the existence and the maintenance of a position of power from which to do the granting." The problem, then, is not only with the specific property or properties that come to be selected as the universal criteria of moral inclusion but also, and perhaps more so, the very act of universalization, which already empowers someone to make these decisions for others.

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, for lack of a better description, “thinking otherwise.” 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 animals, the environment, organizations, and technological artifacts, like machines and robots. In fact, recent debates concerning the moral 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 (Benford and Malartre 2007, 165).

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 machines have rights?” which is largely an ontological query concerned with the prior discovery of intrinsic and morally relevant properties. Instead it is something like “Should machines have rights?” 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 social robotics but moral philosophy itself.

2.1 Relational

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 (2010, 214) 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." As we encounter and interact with others—whether they be other human persons, an animals, the natural environment, or a domestic robot—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 stands 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" (Callicott 1989, 110), instituting what Anne Gerdes (2015), following Coeckelbergh (2012, 49), calls "a relational turn" in ethics.

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 (CSA) studies undertaken by Byron Reeves and Clifford Nass (1996), 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 actual 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" (Reeves and Nass 1996, 22). These results have been verified in two recent studies with robots, one reported in the *International Journal of Social Robotics* (Rosenthal-von der Pütten et al, 2013) where researchers found that human subjects respond emotionally to robots and express empathic concern for machines irrespective of knowledge concerning the properties or inner workings of the mechanism, and another that used physiological evidence, documented by electroencephalography, of the ability of humans to empathize with what appears to be "robot pain" (Suzuki et al, 2015).

2.2 Radically Empirical

Second, 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 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," as Emmanuel Levinas (1987, 56) 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." It is for this reason that Levinasian philosophy focuses attention not on other minds, but on the face of the other. Or as Richard Cohen (2001, 336) 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." [5]

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" 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. Consequently, we can and should perhaps reinterpret Sherry Turkle's criticism of social interactions with robots and read it not as the diagnosis of an ailment but as an accurate description of our very real social situation and circumstance: "I find people willing to seriously consider robots not only as pets but as potential friends, confidants, and even romantic partners. We don't seem to care what their artificial intelligences 'know' or 'understand' of the human moments we might 'share' with them…the performance of connection seems connection enough" (Turkle 2011, 9).

2.3 Altruistic

Finally, 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 (1969, 43), "the strangeness of the Other, his irreducibility to the I, to my thoughts and my possessions, is precisely accomplished as a calling into question of my spontaneity, as ethics." 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"(Singer 1989, 148). 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 for others 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 (2008, 71) 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. 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" (ibid.).

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 (1952, 16) 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." As our world becomes increasingly populated by 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 robot 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. It is, to use the terminology of Thomas Kuhn (1996), widely accepted as "normal science." And a good deal of the current work in moral machines (Allen and Wallach 2009), machine ethics (Anderson and Anderson 2011) and robot ethics (Lin et al. 2012) follows 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, 2) terminological problems with the definition of the morally significant property or properties, 3) epistemological problems with detecting and evaluating these properties in another, and 4) methodological problems caused by the very effort to extend rights to 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 normal science—has limitations and that these limitations are becoming increasingly evident in the face of social robots—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 the question of moral standing that is situated and oriented otherwise. This alternative, which focuses attention on the question of alterity, circumvents many of the problems encountered in the properties approach by arranging for an ethics that is relational, radically empirical, and altruistic. This alternative transaction 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, I would argue, 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 machine 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," Daniel Dennett (1996, vii) writes at the beginning of one of his books, "not a scientist, 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."

For this reason the questions concerning social robotics are not just another problem to be accommodated to and resolved by existing moral theories and procedures. It is instead in the face of increasingly social and interactive machines that moral theory itself also comes to be submitted to a thorough reevaluation and critical questioning. Robo-philosophy, therefore, is not just philosophy applied to the new opportunities and challenges of social robots; it also calls for and requires a thorough reformulation of moral philosophy for and in the face of these other kinds of others.

NOTES

1. The evolution of moral and social philosophy has witnessed the publication of a number of "vindication discourses," which have been pivotal to challenging and eventually altering the scope of moral and legal inclusion. This literature begins with Mary Wollstonecraft's declaration of human rights, *The Vindication of the Rights of Man* (1996), which was initially published in 1790, followed two years later by *The Vindication of the Rights of Women* (Wollstonecraft, 2004). The latter was parodied in Thomas Taylor's (1966/1792) intentionally sarcastic effort at "animal rights" in the *Vindication of the Rights of Butes*. I have sought to contribute to and further develop this tradition with "The Vindication of the Rights of Machines" (Gunkel, 2014), which can be seen as a kind of rejoinder to Taylor's text.

2. Although I mainly focus on the question of rights in this essay, my work (Gunkel 2012) has sought to develop both aspects of the machine question—machine moral agency and machine moral patiency.

3. Although “pain” is not the object of his analysis, the problem of distinguishing between the “real thing” and its mere simulation is illustrated by John Searle's “Chinese Room.” This influential thought experiment, first introduced in 1980 with the essay “Minds, Brains, and Programs” and elaborated in subsequent publications, was offered as an argument against the claims of strong AI. “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” (Searle 1999, 115). 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 (2007, 332) writes, it were possible to design a robot 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.” The problem exhibited by both examples is not simply that there is a difference between simulation and the real thing. The problem is that we also remain persistently unable to distinguish the one from the other in any way that would be considered entirely satisfactory.

4. The word “robot” should not be taken lightly. Like the word “animal,” which Jacques Derrida (2008) argues is a problematic construction in its own right, the collective noun robot is difficult to pin down. This is because, as Kerstin Dautenhahn (2014) correctly points out, “the concept of robot is a moving target, we constantly reinvent what we consider to be 'robot.'" Consequently the term “robot” should be deployed and understood in the way Derrida advises us to deal with “animal.” It is a variable that seeks to collect a number of particular possibilities under its umbrella without ever being able to master completely the entire field.

5. 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*: "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 (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 (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."

REFERENCES

Allen, Colin. 2013. Review of *The Machine Question: Critical Perspectives on AI, Robots, and Ethics. Notre Dame Philosophical Reviews.* https://ndpr.nd.edu/news/37494-the-machine-question-critical-perspectives-on-ai-robots-and-ethics/

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

Atterton, Peter, and Matthew Calarco. 2010. *Radicalizing Levinas*. Albany, NY: State University of New York Press.

Bates, J. 1994. The role of emotion in believable agents. *Communications of the ACM* 37:122–125.

Benford, Gregory, and Elisabeth Malartre. 2007. *Beyond Human: Living with Robots and Cyborgs*. New York: Tom Doherty.

Bentham, Jeremy. 2005. *An Introduction to the Principles of Morals and Legislation*. Ed. J. H. Burns and H. L. Hart. Oxford: Oxford University Press.

Birch, Thomas. 1993. Moral Considerability and Universal Consideration. *Environmental Ethics* 15: 313-332.

Birch, Thomas H. 1995. The incarnation of wilderness: Wilderness areas as prisons. In *Postmodern Environmental Ethics*, ed. Max Oelschlaeger, 137–162. Albany, NY: SUNY Press.

Blumberg, B., P. Todd, and M. Maes. 1996. No bad dogs: Ethological lessons for learning. In *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior* (SAB96), 295–304. Cambridge, MA: MIT Press.

Breazeal, Cynthia, and Rodney Brooks. 2004. Robot Emotion: A Functional Perspective. In *Who Needs Emotions: The Brain Meets the Robot*, ed. J. M. Fellous and M. Arbib, 271–310. Oxford: Oxford University Press.

Brooks, Rodney A. 2002. *Flesh and Machines: How Robots Will Change Us*. New York: Pantheon Books.

Calarco, Matthew. 2008. *Zoographies: The Question of the Animal from Heidegger to Derrida*. New York: Columbia University Press.

Callicott, J. Baird. 1989. *In Defense of the Land Ethic: Essays in Environmental Philosophy*. Albany, NY: State University of New York Press.

Churchland, Paul M. 1999. *Matter and Consciousness* (revised edition). Cambridge, MA: MIT Press.

Coeckelbergh, Mark. 2010. "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." *Ethics and Information Technology* 12:209-221.

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

Cohen, Richard A. 2001. *Ethics, Exegesis, and Philosophy: Interpretation After Levinas*. Cambridge: Cambridge University Press.

Dautenhahn, Kerstin. (2014). Human-robot interaction. In *The Encyclopedia of Human-Computer Interaction,* ed. Soegaard, M. and Dam, R. F. The Interaction Design Foundation, Aarhus, Denmark, 2nd edition. Available online: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction

Dennett, Daniel C. 1996. *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books.

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

Derrida, Jacques. 2008. *The Animal That Therefore I Am*. Ed. Marie-Louise Mallet. Trans. David Wills. New York: Fordham University Press.

Feenberg, Andrew. 1991. *Critical Theory of Technology*. Oxford: Oxford University Press.

Floridi, Luciano. 2013. *The Ethics of Information*. Oxford: Oxford University Press.

Gerdes, Anne. 2015. The Issue of Moral Consideration in Robot Ethics. *ACM SIGCAS Computers & Society* 45(3): 274-280.

Gottlieb, Jeffrey D. 2013. Questions Left Unanswered. *Ethics & Behavior* 23(2): 163-166.

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

Gunkel, David J. 2014. A Vindication of the Rights of Machines. *Philosophy & Technology* 27(1): 113-132.

Güzeldere, Güven. 1997. The many faces of consciousness: A field guide. In *The Nature of Consciousness: Philosophical Debates*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere, 1–68. Cambridge, MA: MIT Press.

Haraway, Donna J. 2008. *When Species Meet*. Minneapolis, MN: University of Minnesota Press.

Heidegger, Martin. 1962. *Being and Time*. Trans. John Macquarrie and Edward Robinson. New York: Harper & Row.

Heidegger, Martin. 1977. *The Question Concerning Technology and Other Essays*. Trans. William Lovitt. New York: Harper & Row.

Himma, Kenneth Einar. 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? E*thics and Information Technology* 11(1): 19–29.

Johnson, Deborah G. 2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology* 8:195–204.

Kant, Immanuel. 1985. *Critique of Practical Reason*. Trans. Lewis White Beck. New York: Macmillan.

Kokoro, L. T. D. 2009. http://www.kokoro-dreams.co.jp/

Kuhn, Thomas S. 1996. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Leopold, Aldo. 1966. *A Sand County Almanac*. Oxford: Oxford University Press.

Levinas, Emmanuel. 1969. *Totality and Infinity: An Essay on Exteriority.* Trans. Alphonso Lingis. Pittsburgh, PA: Duquesne University.

Levinas, Emmanuel. 1987. *Collected Philosophical Papers*. Trans. Alphonso Lingis. Dordrecht: Martinus Nijhoff.

Levy, David. 2005. *Robots Unlimited: Life in a Virtual Age*. Boca Raton, FL: CRC Press.

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

Moore, George E. 2005. *Principia Ethica*. New York: Barnes & Noble Books.

Nealon, Jeffrey. 1998. *Alterity Politics: Ethics and Performative Subjectivity*. Durham, NC: Duke University Press.

O'Regan, Kevin J. 2007. 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*, ed. Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, 332–346. Berlin: Springer-Verlag.

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

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

Rosenthal-von der Pütten, Astrid M., Nicole C. Krämer, Laura Hoffmann, Sabrina Sobieraj and Sabrina C. Eimler. 2013. An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics* 5: 17-34.

Searle, John. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417–457.

Searle, John. 1999. The Chinese room. In *The MIT Encyclopedia of the Cognitive Sciences*, ed. R. A. Wilson and F. Keil, 115–116. Cambridge, MA: MIT Press.

Seibt, Johanna. 2016. Towards an Ontology of Simulated Social Interaction: Varieties of the 'As If' for Robots and Humans. In *Sociality and Normativity for Robots–Philosophical Investigations,* ed. R. Hakli, 1-28. https://www.academia.edu/19415782/Towards_an_Ontology_of_Simulated_Social_Interaction--Varieties_of_the_As_If_for_Robots_and_Humans

Singer, Peter. 1975. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: New York Review of Books.

Singer, Peter. 1989. All animals are equal. In *Animal Rights and Human Obligations*, ed. Tom Regan and Peter Singer, 148–162. New Jersey: Prentice-Hall.

Sparrow, Robert. 2004. The Turing triage test. *Ethics and Information Technology* 6(4):203–213.

Suzuki, Yutaka, Lisa Galli, Ayaka Ikeda, Shoji Itakura and Michiteru Kitazaki. 2015. Measuring empathy for human

and robot hand pain using electroencephalography. *Scientific Reports* 5:15924. http://www.nature.com/articles/srep15924

Taylor, Thomas. 1966. *A Vindication of the Rights of Brutes*. Gainesville, FL: Scholars' Facsimiles & Reprints. Originally published 1792.

Torrance, Steve. 2013. Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Philosophy & Technology* 27(1): 9-29.

Turkle, Sherry. 2012. *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Velmans, Max. 2000. *Understanding Consciousness*. London, UK: Routledge.

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

Wiener, Norbert. 1954. *The Human Use of Human Beings*. New York: Da Capo.

Wollstonecraft, Mary. 1996. *A Vindication of the Rights of Men*. New York: Prometheus Books.

Wollstonecraft, Mary. 2004. *A Vindication of the Rights of Woman*. New York: Penguin Classics.

Žižek, Slavoj. 2008. *For They Know Not What They Do: Enjoyment as a Political Factor*. London: Verso.