A Vindication of the Rights of Machines

David J. Gunkel[1]

Abstract This paper responds to the machine question in the affirmative, arguing that machines, like robots, AI, and other autonomous systems, can no longer be legitimately excluded from moral consideration. The demonstration of this thesis proceeds in three parts. The first and second parts approach the subject by investigating the two constitutive components of the ethical relationship—moral agency and patiency. And in the process, they each demonstrate failure. This occurs not because the machine is somehow unable to achieve what is considered necessary to be considered a moral agent or patient but because the standard characterization of agency and patiency already fail to accommodate not just machines but also those entities who are currently regarded as being moral subjects. The third part responds to this systemic failure by formulating an approach to ethics that is oriented and situated otherwise. This alternative proposes an ethics that is based not on some prior *discovery* concerning the ontological status of others but the product of a *decision* that responds to and is able to be responsible for others and other kinds of otherness.

1. INTRODUCTION

One of the enduring concerns of moral philosophy is determining who or what is deserving of ethical consideration. Although initially limited to "other men," the practice of ethics has developed in such a way that it continually challenges its own restrictions and comes to encompass what had been previously excluded individuals and groups—foreigners, women, animals, and even the environment. "In the history of the United States," Susan Anderson has argued, "gradually more and more beings have been granted the same rights that others possessed and we've become a more ethical society as a result. Ethicists are currently struggling with the question of whether at least some higher animals should have rights, and the status of human fetuses has been debated as well. On the horizon looms the question of whether intelligent machines should have moral standing." [1] The following responds to this final question—what we might call the "machine question" in ethics—in the affirmative, arguing that machines, like robots, AI, and other autonomous systems, can no longer and perhaps never really could be excluded from moral consideration. Toward that end, this paper advances another "vindication discourse," following in a tradition that begins with Mary Wollstonecraft's *A Vindication of the Rights of Men* (1790) succeeded two years later by *A Vindication of the Rights of Woman* and Thomas Taylor's intentionally sarcastic yet remarkably influential response *A Vindication of the Rights of Brutes*.[2]

Although informed by and following in the tradition of these vindication discourses, or what Peter Singer has also called a "liberation movement" [3], the argument presented here will employ something of an unexpected approach and procedure. Arguments for the vindication of the rights of previously excluded others typically proceed by 1) defining or characterizing the criteria for moral considerability or what Thomas Birch calls the conditions for membership in "the club of *consideranda*," [4] and 2) demonstrating that some previously excluded entity or group of entities are in fact capable of achieving a threshold level for inclusion in this community of moral subjects. "The question of considerability has been cast," as Birch explains, "and is still widely understood, in terms of a need for necessary and sufficient conditions which mandate practical respect for whomever or what ever fulfills them." [4] The vindication of the rights of machines, however, will proceed otherwise. Instead of demonstrating that machines or at least one representative machine is able to achieve the necessary and sufficient conditions for moral standing (however that might come to be defined, characterized, and justified) the following both contests this procedure and demonstrates the opposite, showing how the very criteria that have been used to decide the question of moral considerability necessarily fail in the first place. Consequently, the vindication of the rights of machines will not, as one might have initially expected, concern some recent or future success in technology nor will it entail a description of or demonstration with a particular artifact; it will instead investigate a fundamental failure in the procedures of moral philosophy itself—a failure that renders exclusion of the machine both questionable and morally suspect.

2. MORAL AGENCY

Questions concerning moral standing typically begin by addressing agency. The decision to begin with this subject is not accidental, provisional, or capricious. It is dictated and prescribed by the history of moral philosophy, which has traditionally privileged agency and the figure of the moral agent in both theory and practice. As Luciano Floridi explains, moral philosophy, from the time of the ancient Greeks through the modern era and beyond, has been almost exclusively agent-oriented. "Virtue ethics, and Greek

[1] Department of Communication, Northern Illinois University, DeKalb, Illinois 60631, USA Email: dgunkel@niu.edu

[2] What is presented here in the form of a "vindication discourse" is an abbreviated version of an argument that is developed in greater detail and analytical depth in *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. [2]

philosophy more generally," Floridi writes, "concentrates its attention on the moral nature and development of the individual agent who performs the action. It can therefore be properly described as an agent-oriented, 'subjective ethics.'" [5] Modern developments, although shifting the focus somewhat, retain this particular agent-oriented approach. "Developed in a world profoundly different from the small, non-Christian Athens, Utilitarianism, or more generally Consequentialism, Contractualism and Deontologism are the three most well-known theories that concentrate on the moral nature and value of the actions performed by the agent." [5] Although shifting emphasis from the "moral nature and development of the individual agent" to the "moral nature and value" of his or her actions, western philosophy has been, with few exceptions (which we will get to shortly), organized and developed as an agent-oriented endeavor.

When considered from the perspective of the agent, ethics inevitably and unavoidably makes exclusive decisions about *who* is to be included in the community of moral subjects and *what* can be excluded from consideration. The choice of words here is not accidental. As Jacques Derrida points everything turns on and is decided by the difference that separates the "who" from the "what." [6] Moral agency has been customarily restricted to those entities who call themselves and each other "man"—those beings who already give themselves the right to be considered someone who counts as opposed to something that does not. But who counts—who, in effect, gets to be situated under the term "who"—has never been entirely settled, and the historical development of moral philosophy can be interpreted as a progressive unfolding, where what had once been excluded (i.e., women, slaves, people of color, etc.) have slowly and not without considerable struggle and resistance been granted access to the gated community of moral agents and have thereby also come to be someone who counts.

Despite this progress, which is, depending on how one looks at it, either remarkable or insufferably protracted, there remain additional exclusions, most notably non-human animals and machines. Machines in particular have been understood to be mere artifacts that are designed, produced, and employed by human agents for human specified ends. This *instrumentalist* and *anthropocentric* understanding has achieved a remarkable level of acceptance and standard-ization, as is evident by the fact that it has remained in place and largely unchallenged from ancient to postmodern times —from at least Plato's *Phaedrus* to Jean-François Lyotard's *The Postmodern Condition* [7]. Beginning with the animal rights movement, however, there has been considerable pressure to reconsider the ontological assumptions and moral consequences of this legacy of human exceptionalism.

Extending consideration to these other previously marginalized subjects has required a significant reworking of the concept of moral agency, one that is not dependent on genetic make-up, species identification, or some other spurious criteria. As Peter Singer describes it, "the biological facts upon which the boundary of our species is drawn do not have moral significance," and to decide questions of moral agency on this ground "would put us in the same position as racists who give preference to those who are members of their race." [8] For this reason, the question of moral agency has come to be disengaged from identification with the human being and is instead referred to and made dependent upon the generic concept of "personhood." "There appears," G. E. Scott writes, "to be more unanimity as regards the claim that in order for an individual to be a moral agent s/he must possess the relevant features of a person; or, in other words, that being a person is a necessary, if not sufficient, condition for being a moral agent." [9] As promising as this "personist" innovation is, "the category of the person," to reuse terminology borrowed from Marcel Mauss [10], is by no means settled and clearly defined. There is, in fact, little or no agreement concerning what makes someone or something a person and the literature on this subject is littered with different formulations and often incompatible criteria. "One might well hope," Daniel Dennett writes, "that such an important concept, applied and denied so confidently, would have clearly formulatable necessary and sufficient conditions for ascription, but if it does, we have not yet discovered them. In the end there may be none to discover. In the end we may come to realize that the concept person is incoherent and obsolete." [11]

In an effort to contend with, if not resolve this problem, researchers often focus on the one "person making" quality that appears on most, if not all, the lists of "personal properties," whether they include just a couple simple elements [8] or involve numerous "interactive capacities" [12], and that already has traction with practitioners and theorists—consciousness. "Without consciousness," John Locke argued, "there is no person." [13] Or as Kenneth Einar Himma articulates it, "moral agency presupposes consciousness…and that the very concept of agency presupposes that agents are conscious." [14] Formulated in this fashion, moral agency is something that is decided and made dependent on a prior determination of consciousness. If, for example, an animal or a machine can in fact be shown to possess "consciousness," then that entity would, on this account, need to be considered a legitimate moral agent. And not surprisingly, there has been considerable effort in the fields of philosophy, AI, and robotics to address the question of machine moral agency by targeting and examining the question and possibility (or impossibility) of machine consciousness.

This seemingly rational approach, however, runs into considerable ontological and epistemological complications. On the one hand, we do not, it seems, have any widely accepted characterization of "consciousness." The problem, then, is that consciousness, although crucial for deciding who is and who is not a moral agent, is itself a term that is ultimately undecided and considerably equivocal. "The term," as Max Velmans points out, "means many different

things to many different people, and no universally agreed core meaning exists." [15] 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. And 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." [16]

On the other hand, even if it were possible to define consciousness or come to some tentative agreement concerning its necessary and sufficient conditions, we still lack any credible and certain way to determine its actual presence in another. Because consciousness is a property attributed to "other minds," its presence or lack thereof requires access to something that is and remains fundamentally inaccessible. "How does one determine," as Paul Churchland characterizes it, "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?" [17] And the supposed solutions to this "other minds problem," from reworkings and modifications of the Turing Test to functionalist approaches that endeavor to work around this problem altogether [18], only make things more complicated and confused. "There is," as Dennett 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." [11] 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 empirical science, strictly speaking.[3] 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 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)." [19]

The question of machine moral agency, therefore, turns out to be anything but simple or definitive. This is not, it is important to note, because machines are somehow unable to be moral agents. It is rather a product of the fact that the term "moral agent," for all its importance and argumentative expediency, has been and remains an ambiguous, indeterminate, and rather noisy concept. What the consideration of machine moral agency demonstrates, therefore, is something that may not have been anticipated or sought. What is discovered in the process of pursuing this line of inquiry is not a satisfactory answer to the question whether machines are able to be moral agents or not. In fact, that question remains open and unanswered. What has been ascertained is that the concept of moral agency is already vague and imprecise such that it is (if applied strictly and rigorously) uncertain whether we—whoever this "we" includes—are in fact moral agents.

What the machine question demonstrates, therefore, is that moral agency, the issue that had been assumed to be the "correct" place to begin, turns out to be inconclusive. Although this could be regarded as a "failure," it is a particularly instructive failing. What is learned from this failure—assuming we continue to use this obviously "negative" word—is that moral agency is not necessarily some property that can be definitively ascertained or discovered in others prior to and in advance of their moral consideration. Instead moral standing may be something (perhaps what Kay Foerst has called a dynamic and socially constructed "honorarium" [20]) that comes to be conferred and assigned to others in the process of our interactions and relationships with them. But then the deciding issue will no longer be one of agency; it will be a matter of *patiency*.

3. MORAL PATIENCY

Moral patiency looks at the ethical relationship from the other side. It is concerned not with determining the moral character of the agent or weighing the ethical significance of his/her/its actions but with the victim, recipient, or receiver of such action. This approach is, as Mane Hajdin [21], Luciano Floridi [5], and others have recognized, a significant alteration in procedure and a "non-standard" way to approach the question of moral rights and responsibilities. The model for this kind of transaction can be found in animal rights philosophy. Whereas agent-oriented ethics have been concerned with determining whether someone is or is not a legitimate moral subject with rights and responsibilities, animal rights philosophy begins with an entirely different question—"Can they suffer?" [22]

This seemingly simple and direct inquiry introduces what turns out to be a major paradigm shift in the basic structure and procedures of moral thinking. On the one hand, it challenges the anthropocentric tradition in ethics by questioning the often unexamined privilege human beings have granted themselves. In effect, it institutes something

[3] Attempts to resolve this problem often take the form of a pseudo-science called *physiognomy*, which endeavors to infer an entity's internal states of mind from the observation of its external expressions and behavior.

like a Copernican revolution in moral philosophy. Just as Copernicus challenged the geocentric model of the cosmos and in the process undermined many of the presumptions of human exceptionalism, animal rights philosophy contests the established Ptolemaic system of ethics, deposing the anthropocentric privilege that had traditionally organized the moral universe. On the other hand, the effect of this fundamental shift in focus means that the one time closed field of ethics can be opened up to other kinds of non-human animals. In other words, who counts as morally significant are not just other "men" but all kinds of entities that had previously been marginalized and situated outside the gates of the moral community. "If a being suffers," Peter Singer writes, "there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that its suffering be counted equally with the like suffering of any other being." [23]

Initially there seems to be good reasons and opportunities for extending this innovation to machines, or at least some species of machines. [24] This is because the animal and the machine, beginning with the work of René Descartes, share a common ontological status and position. For Descartes, the human being was considered the sole creature capable of rational thought—the one entity able to say, and be certain in its saying, *cogito ergo sum*. Following from this, he had concluded that other animals not only lacked reason but were nothing more than mindless automata that, like clockwork mechanisms, simply followed predetermined instructions programmed in the disposition of their various parts or organs. Conceptualized in this fashion, the animal and the machine, or what Descartes identified with the hybrid, hyphenated term *bête-machine*, were effectively indistinguishable and ontologically the same. "If any such machine," Descartes wrote, "had the organs and outward shape of a monkey or of some other animal that lacks reason, we should have no means of knowing that they did not possess entirely the same nature as these animals." [25]

Despite this fundamental and apparently irreducible similitude, only one of the pair has been considered a legitimate subject of moral concern. Even though the fate of the machine, from Descartes forward was intimately coupled with that of the animal, only the animal (and only some animals, at that) has qualified for any level of ethical consideration. And this exclusivity has been asserted and justified on the grounds that the machine, unlike the animal, does not experience either pleasure or pain. Although this conclusion appears to be rather reasonable and intuitive, it fails for a number of reasons.

First, it has been practically disputed by the construction of various mechanisms that now appear to suffer or at least provide external evidence of something that looks like pain. As Derrida recognized, "Descartes already spoke, as if by chance, of a machine that simulates the living animal so well that it 'cries out that you are hurting it.'" [26] This comment, which appears in a brief parenthetical aside in Descartes' *Discourse on Method*, had been deployed in the course of an argument that sought to differentiate human beings from the animal by associating the latter with mere mechanisms. But the comment can, in light of the procedures and protocols of animal ethics, be read otherwise. That is, if it were indeed possible to construct a machine that did exactly what Descartes had postulated, that is, "cry out that you are hurting it," would we not also be obligated to conclude that such a mechanism was capable of experiencing pain? This is, it is important to note, not just a theoretical point or speculative thought experiment. Engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses [27] [28] [29], like the dental-training robot Simroid "who" cries out in pain when students "hurt" it [30], but also systems capable of evidencing behaviors that look a lot like what we usually call pleasure and pain.

Second it can be contested on epistemologically grounds insofar as suffering or the experience of pain is still unable to get around or resolve the problem of other minds. How, for example, can one know that an animal or even another person actually suffers? How is it possible to access and evaluate the suffering that is experienced by another? "Modern philosophy," Matthew Calarco writes, "true to its Cartesian and scientific aspirations, is interested in the indubitable rather than the undeniable. Philosophers want proof that animals actually suffer, that animals are aware of their suffering, and they require an argument for why animal suffering should count on equal par with human suffering." [31] But such indubitable and certain knowledge, as explained by Marian S. Dawkins, appears to be unattainable:

> At first sight, 'suffering' and 'scientific' are not terms that can or should be considered together. When applied to ourselves, 'suffering' refers to the subjective experience of unpleasant emotions such as fear, pain and frustration that are private and known only to the person experiencing them. To use the term in relation to non-human animals, therefore, is to make the assumption that they too have subjective experiences that are private to them and therefore unknowable by us. 'Scientific' on the other hand, means the acquisition of knowledge through the testing of hypotheses using publicly observable events. The problem is that we know so little about human consciousness that we do not know what publicly observable events to look for in ourselves, let alone other species, to ascertain whether they are subjectively experiencing anything like our suffering. The scientific study of animal suffering would, therefore, seem to rest on an inherent contradiction: it requires the testing of the untestable. [32]

Because suffering is understood to be a subjective and private experience, there is no way to know, with any

certainty or credible empirical method, how another entity experiences unpleasant sensations such as fear, pain, or frustration. For this reason, it appears that the suffering of another (especially an animal) remains fundamentally inaccessible and unknowable. As Singer [23] readily admits, "we cannot directly experience anyone else's pain, whether that 'anyone' is our best friend or a stray dog. Pain is a state of consciousness, a 'mental event,' and as such it can never be observed." The machine question, therefore, leads to an outcome that was not necessarily anticipated. The basic problem is not whether the question "can they suffer?" applies to machines but whether anything that appears to suffer—human, animal, plant, or machine—actually does so at all.

Third, and to make matters even more complicated, we may not even know what "pain" and "the experience of pain" is in the first place. This point is something that is taken up and demonstrated by Daniel Dennett's "Why You Can't Make a Computer That Feels Pain." In this provocatively titled essay, originally published decades before the debut of even a rudimentary working prototype, 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." [11] At the end of what turns out to be a rather protracted and detailed consideration of the problem, he 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, nor does it offer any kind of support for the advocates of moral exceptionalism. 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. The best we are able to do, as Dennett illustrates, is account for the various "causes and effects of pain," but "pain itself does not appear." [11] What is demonstrated, 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." [11] Although Bentham's question "Can they suffer?" [22] may have radically reoriented the direction of moral philosophy, the fact remains that "pain" and "suffering" are just as nebulous and difficult to define and locate as the concepts they were intended to replace.

Finally, all this talk about the possibility of engineering pain or suffering in a machine entails its own particular moral dilemma. "If (ro)bots might one day be capable of experiencing pain and other affective states," Wendell Wallach and Colin Allen write, "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 and should it be prohibited?" [18] If it were in fact possible to construct a machine that "feels pain" (however defined and instantiated) in order to demonstrate the limits of moral patiency, then doing so might be ethically suspect insofar as in constructing such a mechanism we do not do everything in our power to minimize its suffering. Consequently, moral philosophers and robotics engineers find themselves in a curious and not entirely comfortable situation. One needs to be able to construct such a mechanism in order to demonstrate moral patiency and the moral standing of machines; but doing so would be, on that account, already to engage in an act that could potentially be considered immoral. Or to put it another way, the demonstration of machine moral patiency might itself be something that is quite painful for others.

For these reasons, approaching the machine question from the perspective of moral patiency also encounters fundamental difficulties. Despite initial promises, we cannot, it seems, make a credible case for or against the moral standing of the machine by simply following the patient-oriented approach modeled by animal rights philosophy. In fact, trying to do so produces some rather unexpected results. In particular, extending these innovations does not provide definitive proof that the machine either can be or is not able to be a similarly constructed moral patient. Instead doing so demonstrates how the "animal question"—the question that has in effect revolutionized ethics in the later half of the 20th century—might already be misguided and prejudicial. Although it was not necessarily designed to work in this fashion, "A Vindication of the Rights of Machines" achieves something similar to what Thomas Taylor had wanted for his *A Vindication of the Rights of Brutes*. Taylor, who wrote and distributed this pamphlet under the protection of anonymity, originally composed his essay as a means by which to parody and undermine the arguments that had been advanced in Wollstonecraft's *A Vindication of the Rights of Woman*. [23] Taylor's text, in other words, was initially offered as a kind of *reductio ad absurdum* designed to exhibit what he perceived to be the conceptual failings of Wollstonecraft's proto-feminist manifesto. Following suit, "A Vindication of the Rights of Machines" appears to have the effect of questioning and even destabilizing what had been achieved with animal rights philosophy. But as was the case with the consideration of moral agency, this negative outcome is informative and telling. In particular, it indicates to what extent this apparent revolution in moral thinking is, for all its insight and promise, still beset with fundamental problems that proceed not so much from the ontological condition of these other, previously excluded entities but from systemic problems in the very structure and protocols of moral reasoning.

4. ULTERIOR MORALS

"Every philosophy," Silvo Benso writes in a comprehensive gesture that performs precisely what it seeks to address, "is a quest for wholeness." [33] This objective, she argues, has been typically targeted in one of two ways. "Traditional Western thought has pursued wholeness by means of reduction, integration, systematization of all its parts. Totality has replaced wholeness, and the result is totalitarianism from which what is truly other escapes, revealing the deficiencies and fallacies of the attempted system." [33] This is precisely the kind of violent philosophizing that Emmanuel Levinas identifies under the term "totality," and which includes, for him at least, the big landmark figures like Plato, Kant, and Heidegger. [34] The alternative to this totalizing approach is a philosophy that is oriented otherwise, like that proposed and developed by Singer, Birch, Levinas, and others. This other approach, however, "must do so by moving not from the same, but from the other, and not only the Other, but also the other of the Other, and, if that is the case, the other of the other of the Other. In this *must*, it must also be aware of the inescapable injustice embedded in any formulation of the other." [33] What is interesting about these two strategies is not what makes them different from one another or how they articulate approaches that proceeds from what appears to be opposite ends of the spectrum. What is interesting is what they agree upon and hold in common in order to be situated as different from and in opposition to each other in the first place. Whether taking the form of autology or some kind of heterology, "they both share the same claim to inclusiveness" [33], and that is the problem.

When it comes to including previously excluded subjects, then, moral philosophy appears to be caught between a proverbial rock and a hard place. On the one hand, the same has never been inclusive enough to adequately accommodate others. The machine in particular is already and from the very beginning situated outside ethics. It is, irrespective of the different philosophical perspectives that come to be mobilized, typically regarded as neither a legitimate moral agent nor patient. It has been and continues to be widely understood as nothing more than an instrument to be employed more or less effectively by human beings and, for this reason, is always and already located in excess of moral considerability or to use that distinct Nietzschean characterization, "beyond good and evil." [35] Technology, as Lyotard reminds us, is only a matter of efficiency. Technical devices do not participate in the big questions of metaphysics, aesthetics, or ethics. [7] They are nothing more than contrivances or extensions of human agency, used more or less responsibly by human agents with the outcome effecting other human patients. Although other kinds of previously marginalized others—animals, the environment, and even corporations—have been slowly and not without considerable struggle granted some level of membership in the community of moral subjects, the machine is and remains on the periphery. "We have never," as J. Storrs Hall points out, "considered ourselves to have 'moral' duties to our machines, or them to us." [36]

On the other hand, alternatives to this tradition, like the patient-oriented approach of animal rights philosophy, have never been different enough. Although a concern with and for others promised to radicalize the procedures of moral reasoning, ethics has not been suitably different. Many of the so-called alternatives, those philosophies that purport to be interested in and oriented otherwise, have typically excluded the machine from what is considered Other. Technological devices certainly have an interface but they do not, as Levinas would have it, possess a face or confront us in a face-to-face encounter that would call for and would be called ethics. [34] This exclusivity is not simply the last socially accepted prejudice or what Singer calls "the last remaining form of discrimination" [3], which may be identified as such only from a perspective that is already open to the possibility of some future inclusion and accommodation. The marginalization of the machine appears to be much more complete and pervasive. In fact, the machine does not constitute just one more form of difference that would be included at some future time. It comprises the very mechanism of exclusion. "In the eyes of many philosophers," Dennett writes, "the old question of whether determinism (or indeterminism) is incompatible with moral responsibility has been superseded by the hypothesis that mechanism may well be." [11] Consequently, whenever a philosophy endeavors to make a decision, to demarcate and draw the line separating "us" from "them," or to differentiate who does and what does not have moral standing, it inevitably fabricates machines. When Tom Regan, for instance, sought to distinguish which higher-order animals qualify for moral consideration as opposed to those lower-order entities that do not, he marginalizes the latter by characterizing them as mere machines. [37]

For these reasons, the machine does not constitute one more historically marginalized other that would need to be granted admission to the class of moral *consideranda*. In fact, it now appears that the machine is unable to achieve what is considered to be necessary for either moral agency or patiency. But this inability is not, we can say following the argumentative strategy of Dennett's "Why you Cannot Make a Computer that Feels Pain" [11], a product of some inherent or essential deficiency with the machine. Instead it is a result of the fact that both agency and patiency already lack clearly defined necessary and sufficient conditions. "A Vindication of the Rights of Machines," then, does not end by accumulating evidence or arguments in favor of permitting one more entity entry into the community of moral subjects; it concludes by critically questioning the very protocols of inclusion/exclusion that have organized and structured moral philosophy from the beginning.

What this means for ethics is that Descartes—that thinker who had been regarded as the "bad guy" of modern philosophy by Regan [37] and others [38]—may have

actually gotten it right despite himself and our usual (mis)interpretations of his work. In the *Discourse on the Method*, something of a philosophical autobiography, Descartes famously endeavored to tear down to its foundations every truth that he had come to accept or had taken for granted. This approach, which will in the *Meditations* come to be called "the method of doubt," targets everything, including the accepted truths of ethics. With Descartes, then, one thing is certain, he did not want to be nor would he tolerate being duped. However, pursuing and maintaining this extreme form of critical inquiry that does not respect any pre-established boundaries has very real practical expenses and implications. For this reason, Descartes decides to adopt a "provisional moral code," something of a temporary but relatively stable structure that would support and shelter him as he engaged in this thorough questioning of everything and anything.

> Now, before starting to rebuild your house, it is not enough simply to pull it down, to make provision for materials and architects (or else train yourself in architecture), and to have carefully drawn up the plans; you must also provide yourself with some other place where you can live comfortably while building is in progress. Likewise, lest I should remain indecisive in my actions while reason obliged me to be so in my judgments, and in order to live as happily as I could during this time, I formed for myself a provisional moral code consisting of just three or four maxims, which I should like to tell you about. [25]

Understood and formulated as "provisional," it might be assumed that this protocol would, at some future time, be replaced by something more certain and permanent. But Descartes, for whatever reason, never explicitly returns to it in order to finalize things. This is, despite initial appearances, not a deficiency, failure, or oversight. It may, in fact, be the truth of the matter. Namely that, as Slavoj Žižek describes it, "all morality we adopt is provisory." [39] In this case, then, what would have customarily been considered "failure," that is, the lack of ever achieving the *terra firma* of moral certitude, is reconceived of as a kind of success and advancement. Consequently, "failure," Žižek argues, "is no longer perceived as opposed to success, since success itself can consist only in heroically assuming the full dimension of failure itself, 'repeating' failure as 'one's own.'" [39] In other words, the provisory nature of ethics is not a failure as opposed to some other presumed outcome that would be called "success." It is only by assuming and affirming this supposed "failure" that what is called ethics will have succeeded.

Ethics, conceived of in this fashion, is not determined by a prior ontological discovery concerning the essential capabilities or internal operations of others. It is rather a decision—literally a cut that institutes difference and that makes a difference by dividing between *who* is considered to be morally significant and *what* is not. Consequently, "moral consideration is," as Mark Coeckelbergh describes it, "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." [40] This is the reason why, as Levinas claims, "morality is first philosophy" ("first" in terms of both sequence and status) and that moral decision making precedes ontological knowledge. [34][4] What this means, in the final analysis, is that we—we who already occupy a privileged position within the community of moral subjects—are responsible for determining the proper scope and boundaries of moral responsibility, for instituting these decisions in everyday practices, and for evaluating their results and outcomes. Although we have often sought to deflect these decisions and responsibilities elsewhere, typically into the heavens but also onto other terrestrial authorities, in order to validate and/or to avoid having to take responsibility for them, we are, in the final analysis, the sole responsible party. We are, in other words, not just responsible for acting responsibly in accordance with ethics; we are responsible for ethics. The vindication of the rights of machines, therefore, is not simply a matter of extending moral consideration to one more historically excluded other. The question concerning the "rights of machines" makes a fundamental claim on ethics, requiring us to rethink the system of moral considerability all the way down.

REFERENCES

[1] S. Anderson. Asimov's "Three Laws of Robotics" and Machine Metaethics. *AI & Society* 22(4): 477–493 (2008)

[2] D. Gunkel. *The Machine Question: Critical Perspectives on AI, Robots and Ethics.* MIT Press, USA (2012).

[3] P. Singer. All Animals are Equal. In *Animal Rights and Human Obligations*. Tom Regan and Peter Singer (Eds), pp. 148–162. Prentice Hall, USA (1989).

[4] T. Birch. Moral Considerability and Universal Consideration. *Environmental Ethics* 15: 313–332 (1993).

[5] L. Floridi. Information Ethics: On the Philosophical Foundation of Computer Ethics. *Ethics and Information Technology* 1(1): 37–56 (1999).

[6] J. Derrida. *Paper Machine*. Rachel Bowlby (Trans.). Stanford University Press, USA (2005).

[7] J. Lyotard. *The Postmodern Condition: A Report on Knowledge*. G. Bennington and B. Massumi (Trans.). University of Minnesota Press, USA (1984).

[8] P. Singer. *Practical Ethics*. Cambridge University Press, UK (1999).

[9] G. E. Scott. *Moral Personhood: An Essay in the Philosophy of Moral Psychology*. SUNY Press, USA (1990).

[10] M. Mauss. A Category of the Human Mind: The Notion of Person; The Notion of Self. W. D. Halls (Trans.). In *The Category of the Person*, M. Carrithers, S. Collins, and S. Lukes (Eds.), pp. 1–25. Cambridge University Press, UK (1985).

[11] D. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, USA (1998).

[4] If there is a philosophical tradition that is explicitly dedicated to pursuing this procedural inversion, it is arguably German Idealism in general and Kant in particular, all of which assert, as Žižek characterizes it, "the primacy of practical over theoretical reason." [41]

[12] C. Smith. *What Is a Person? Rethinking Humanity, Social Life, and the Moral Good from the Person Up*. University of Chicago Press, USA (2010).
[13] J. Locke. *An Essay Concerning Human Understanding*. Hackett, USA (1996).
[14] K. E. Himma. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent? *Ethics and Information Technology* 11(1):19–29 (2009).
[15] M. Velmans. *Understanding Consciousness*. Routledge, UK (2000).
[16] G. Güzeldere. The Many Faces of Consciousness: A Field Guide. In The Nature *of Consciousness: Philosophical Debates*, N. Block, O. Flanagan, and G. Güzeldere (Eds.), pp. 1–68. MIT Press, USA (1997).
[17] P. M. Churchland. *Matter and Consciousness*, rev. ed. Cambridge, MIT Press, USA (1999).
[18] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, UK (2009).
[19] R. Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. Viking, USA (2005).
[20] G. Benford and E. Malartre. 2007. *Beyond Human: Living with Robots and Cyborgs*. Tom Doherty, USA (2007).
[21] M. Hajdin. *The Boundaries of Moral Discourse*. Loyola University Press (1994).
[22] J. Bentham. *An Introduction to the Principles of Morals and Legislation*. J. H. Burns and H. L. Hart (Eds.). Oxford University Press, UK (2005).
[23] P. Singer. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York Review of Books, USA (1975).
[24] D. J. Gunkel. *Thinking Otherwise: Philosophy, Communication, Technology*. Purdue University Press, USA (2007).
[25] R. Descartes. *Selected Philosophical Writings*. J. Cottingham, R. Stoothoff, and D. Murdoch (Trans.). Cambridge University Press, UK (1988).
[26] J. Derrida. *The Animal That Therefore I Am*. M. Mallet (Ed.) and David Wills (Trans.). Fordham University Press, USA (2008).
[27] J. Bates. The Role of Emotion in Believable Agents. *Communications of the ACM* 37: 122–125 (1994).
[28] 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* (SAB96), pp. 295–304. MIT Press, USA (1996).
[29] C. Breazeal and R. Brooks. Robot Emotion: A Functional Perspective. In *Who Needs Emotions: The Brain Meets the Robot*, J. M. Fellous and M. Arbib (Eds.), pp. 271–310. Oxford University Press, UK (2004).
[30] Kokoro, L. T. D. http://www.kokoro-dreams.co.jp/ (2009).
[31] M. Calarco. *Zoographies: The Question of the Animal from Heidegger to Derrida*. Columbia University Press, USA (2008)
[32] M. S. Dawkins. The Science of Animal Suffering. *Ethology* 114(10): 937–945 (2008).
[33] S. Benso. *The Face of Things: A Different Side of Ethics*. SUNY Press, USA (2000).
[34] E. Levinas. *Totality and Infinity: An Essay on Exteriority.* A. Lingis (Trans.). Duquesne University Press, USA (1969).
[35] F. Nietzsche. *Beyond Good and Evil.* W. Kaufmann (Trans.). Vintage Books, USA (1966).
[36] J. S. Hall. Ethics for Machines. http://www.kurzweilai.net/ethics-for-machines. (5 July 2001).
[37] T. Regan. *The Case for Animal Rights*. University of California Press, USA (1983).
[38] A. Lippit. *Electric Animal: Toward a Rhetoric of Wildlife*. University of Minnesota Press, USA (2000).
[39] S. Žižek. *The Parallax View*. MIT Press, USA (2006).
[40] M. Coeckelbergh. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology*, 12(3): 209-221 (2010).
[41] S. Žižek. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*. Verso, USA (2012).