THE MACHINE QUESTION

CRITICAL PERSPECTIVES ON AI, ROBOTS, AND ETHICS

DAVID J. GUNKEL

The Machine Question
Critical Perspectives on AI, Robots, and Ethics

David J. Gunkel

The MIT Press
Cambridge, Massachusetts
London, England

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email special_sales@mitpress.mit.edu or write to Special Sales Department, The MIT Press, 55 Hayward Street, Cambridge, MA 02142.

This book was set in Stone Sans and Stone Serif by Toppan Best-set Premedia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Gunkel, David J.
The machine question : critical perspectives on AI, robots, and ethics /
David J. Gunkel.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 978-0-262-01743-5 (hardcover : alk. paper)
1. Robotics—Human factors. 2. Robotics—Moral and ethical aspects. 3. Robotics—Philosophy. 4. Artificial intelligence. I. Title.
TJ211.49.G86 2012
174'.9629892—dc23

2011045428

10 9 8 7 6 5 4 3 2 1

Contents

Introduction

One of the enduring concerns of moral philosophy is deciding 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. Currently, we stand on the verge of another fundamental challenge to moral thinking. This challenge comes from the autonomous, intelligent machines of our own making, and it puts in question many deep-seated assumptions about who or what constitutes a moral subject. The way we address and respond to this challenge will have a profound effect on how we understand ourselves, our place in the world, and our responsibilities to the other entities encountered here.

Take for example one of the quintessential illustrations of both the promise and peril of autonomous machine decision making, Stanley Kubrick's *2001: A Space Odyssey* (1968). In this popular science fiction film, the HAL 9000 computer endeavors to protect the integrity of a deep-space mission to Jupiter by ending the life of the spacecraft's human crew. In response to this action, the remaining human occupant of the spacecraft terminates HAL by shutting down the computer's higher cognitive functions, effectively killing this artificially intelligent machine. The scenario obviously makes for compelling cinematic drama, but it also illustrates a number of intriguing and important philosophical problems: Can machines be held responsible for actions that affect human beings? What limitations, if any, should guide autonomous decision making by artificial intelligence systems, computers, or robots? Is it possible to program such mechanisms

with an appropriate sense of right and wrong? What moral responsibilities would these machines have to us, and what responsibilities might we have to such ethically minded machines?

Although initially presented in science fiction, these questions are increasingly becoming science fact. Researchers working in the fields of artificial intelligence (AI), information and communication technology (ICT), and robotics are beginning to talk quite seriously about ethics. In particular, they are interested in what is now called the ethically programmed machine and the moral standing of artificial autonomous agents. In the past several years, for instance, there has been a noticeable increase in the number of dedicated conferences, symposia, and workshops with provocative titles like "Machine Ethics," "EthicALife," "AI, Ethics, and (Quasi)Human Rights," and "Roboethics"; scholarly articles and books addressing this subject matter like Luciano Floridi's "Information Ethics" (1999), J. Storrs Hall's "Ethics for Machines" (2001), Anderson et al.'s "Toward Machine Ethics" (2004), and Wendell Wallach and Colin Allen's *Moral Machines* (2009); and even publicly funded initiatives like South Korea's Robot Ethics Charter (see Lovgren 2007), which is designed to anticipate potential problems with autonomous machines and to prevent human abuse of robots, and Japan's Ministry of Economy, Trade and Industry, which is purportedly working on a code of behavior for robots, especially those employed in the elder care industry (see Christensen 2006).

Before this new development in moral thinking advances too far, we should take the time to ask some fundamental philosophical questions. Namely, what kind of moral claim might such mechanisms have? What are the philosophical grounds for such a claim? And what would it mean to articulate and practice an ethics of this subject? *The Machine Question* seeks to address, evaluate, and respond to these queries. In doing so, it is designed to have a fundamental and transformative effect on both the current state and future possibilities of moral philosophy, altering not so much the rules of the game but questioning who or what gets to participate.

The Machine Question

If there is a "bad guy" in contemporary philosophy, that title arguably belongs to René Descartes. This is not because Descartes was a particularly

bad individual or did anything that would be considered morally suspect. Quite the contrary. It is simply because he, in the course of developing his particular brand of modern philosophy, came to associate the animal with the machine, introducing an influential concept—the doctrine of the *bête-machine* or *animal-machine*. "Perhaps the most notorious of the dualistic thinkers," Akira Mizuta Lippit (2000, 33) writes, "Descartes has come to stand for the insistent segregation of the human and animal worlds in philosophy. Likening animals to automata, Descartes argues in the 1637 *Discourse on the Method* that not only 'do the beasts have less reason than men, but they have no reason at all.'" 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 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" (Descartes 1988, 44). Beginning with Descartes, then, the animal and machine share a common form of alterity that situates them as completely different from and distinctly other than human. Despite pursuing of a method of doubt that, as Jacques Derrida (2008, 75) describes it, reaches "a level of hyperbole," Descartes "never doubted that the animal was only a machine."

Following this decision, animals have not traditionally been considered a legitimate subject of moral concern. Determined to be mere mechanisms, they are simply instruments to be used more or less effectively by human beings, who are typically the only things that matter. When Kant (1985), for instance, defined morality as involving the rational determination of the will, the animal, which does not by definition possess reason, is immediately and categorically excluded. The practical employment of reason does not concern the animal, and, when Kant does make mention of animality (*Tierheit*), he does so only in order to use it as a foil by which to define the limits of humanity proper. Theodore Adorno, as Derrida points out in the final essay of *Paper Machine*, takes the interpretation one step

further, arguing that Kant not only excluded animality from moral consideration but held everything associated with the animal in contempt: "He [Adorno] particularly blames Kant, whom he respects too much from another point of view, for not giving any place in his concept of dignity (*Würde*) and the 'autonomy' of man to any compassion (*Mitleid*) between man and the animal. Nothing is more odious (*verhasster*) to Kantian man, says Adorno, than remembering a resemblance or affinity between man and animal (*die Erinnerung an die Tierähnlichkeit des Menschen*). The Kantian feels only hate for human animality" (Derrida 2005, 180). The same ethical redlining was instituted and supported in the analytic tradition. According to Tom Regan, this is immediately apparent in the seminal work of analytical ethics. "It was in 1903 when analytic philosophy's patron saint, George Edward Moore, published his classic, *Principia Ethica*. You can read every word in it. You can read between every line of it. Look where you will, you will not find the slightest hint of attention to 'the animal question.' Natural and nonnatural properties, yes. Definitions and analyses, yes. The open-question argument and the method of isolation, yes. But so much as a word about nonhuman animals? No. Serious moral philosophy, of the analytic variety, back then did not traffic with such ideas" (Regan 1999, xii).

It is only recently that the discipline of philosophy has begun to approach the animal as a legitimate subject of moral consideration. Regan identifies the turning point in a single work: "In 1971, three Oxford philosophers—Roslind and Stanley Godlovitch, and John Harris—published *Animals, Men and Morals*. The volume marked the first time philosophers had collaborated to craft a book that dealt with the moral status of nonhuman animals" (Regan 1999, xi). According to Regan, this particular publication is not only credited with introducing what is now called the "animal question," but launched an entire subdiscipline of moral philosophy where the animal is considered to be a legitimate subject of ethical inquiry. Currently, philosophers of both the analytic and continental varieties find reason to be concerned with animals, and there is a growing body of research addressing issues like the ethical treatment of animals, animal rights, and environmental ethics.

What is remarkable about this development is that at a time when this form of nonhuman otherness is increasingly recognized as a legitimate moral subject, its other, the machine, remains conspicuously absent and

marginalized. Despite all the ink that has been spilled on the animal question, little or nothing has been written about the machine. One could, in fact, redeploy Regan's critique of G. E. Moore's *Principia Ethica* and apply it, with a high degree of accuracy, to any work purporting to address the animal question: "You can read every word in it. You can read between every line of it. Look where you will, you will not find the slightest hint of attention to 'the machine question.'" Even though the fate of the machine, from Descartes forward, was intimately coupled with that of the animal, only one of the pair has qualified for any level of ethical consideration. "We have," in the words of J. Storrs Hall (2001), "never considered ourselves to have 'moral' duties to our machines, or them to us." The machine question, therefore, is the other side of the question of the animal. In effect, it asks about the other that remains outside and marginalized by contemporary philosophy's recent concern for and interest in others.

Structure and Approach

Formulated as an ethical matter, the machine question will involve two constitutive components. "Moral situations," as Luciano Floridi and J. W. Sanders (2004, 349–350) point out, "commonly involve agents and patients. Let us define the class *A* of moral *agents* as the class of all entities that can in principle qualify as sources of moral action, and the class *P* of moral *patients* as the class of all entities that can in principle qualify as receivers of moral action." According to the analysis provided by Floridi and Sanders (2004, 350), there "can be five logical relations between *A* and *P*." Of these five, three are immediately set aside and excluded from further consideration. This includes situations where *A* and *P* are disjoint and not at all related, situations where *P* is a subset of *A*, and situations where *A* and *P* intersect. The first formulation is excluded from serious consideration because it is determined to be "utterly unrealistic." The other two are set aside mainly because they require a "pure agent"—"a kind of supernatural entity that, like Aristotle's God, affects the world but can never be affected by it" (Floridi and Sanders 2004, 377).[1] "Not surprisingly," Floridi and Sanders (2004, 377) conclude, "most macroethics have kept away from these supernatural speculations and implicitly adopted or even explicitly argued for one of the two remaining alternatives."

> Alternative (1) maintains that all entities that qualify as moral agents also qualify as moral patients and vice versa. It corresponds to a rather intuitive position, according to which the agent/inquirer plays the role of the moral protagonist, and is one of the most popular views in the history of ethics, shared for example by many Christian Ethicists in general and by Kant in particular. We refer to it as the standard position. Alternative (2) holds that all entities that qualify as moral agents also qualify as moral patients but not vice versa. Many entities, most notably animals, seem to qualify as moral patients, even if they are in principle excluded from playing the role of moral agents. This post-environmentalist approach requires a change in perspective, from agent orientation to patient orientation. In view of the previous label, we refer to it as non-standard. (Floridi and Sanders 2004, 350)

Following this arrangement, which is not something that is necessarily unique to Floridi and Sanders's work (see Miller and Williams 1983; Regan 1983; McPherson 1984; Hajdin 1994; Miller 1994), the machine question will be formulated and pursued from both an agent-oriented and patient-oriented perspective.

The investigation begins in chapter 1 by addressing the question of machine moral agency. That is, it commences by asking whether and to what extent machines of various designs and functions might be considered a legitimate moral agent that could be held responsible and accountable for decisions and actions. Clearly, this mode of inquiry already represents a major shift in thinking about technology and the technological artifact. For most if not all of Western intellectual history, technology has been explained and conceptualized as a tool or instrument to be used more or less effectively by human agents. As such, technology itself is neither good nor bad, it is just a more or less convenient or effective means to an end. This "instrumental and anthropological definition of technology," as Martin Heidegger (1977a, 5) called it, is not only influential but is considered to be axiomatic. "Who would," Heidegger asks rhetorically, "ever deny that it is correct? It is in obvious conformity with what we are envisioning when we talk about technology. The instrumental definition of technology is indeed so uncannily correct that it even holds for modern technology, of which, in other respects, we maintain with some justification that it is, in contrast to the older handwork technology, something completely different and therefore new. . . . But this much remains correct: modern technology too is a means to an end" (ibid.).

In asking whether technological artifacts like computers, artificial intelligence, or robots can be considered moral agents, chapter 1 directly and

quite deliberately challenges this "uncannily correct" characterization of technology. To put it in a kind of shorthand or caricature, this part of the investigation asks whether and to what extent the standard dodge of the customer service representative—"I'm sorry, sir, it's not me. It's the computer"—might cease being just another lame excuse and become a situation of legitimate machine responsibility. This fundamental reconfiguration of the question concerning technology will turn out to be no small matter. It will, in fact, have significant consequences for the way we understand technological artifacts, human users, and the presumed limits of moral responsibility.

In chapter 2, the second part of the investigation approaches the machine question from the other side, asking to what extent machines might constitute an Other in situations of moral concern and decision making. It, therefore, takes up the question of whether machines are capable of occupying the position of a moral patient "who" has a legitimate claim to certain rights that would need to be respected and taken into account. In fact, the suspension of the word "who" in quotation marks indicates what is at stake in this matter. "Who" already accords someone or something the status of an Other in social relationships. Typically "who" refers to other human beings—other "persons" (another term that will need to be thoroughly investigated) who like ourselves are due moral respect. In contrast, things, whether they are nonhuman animals, various living and nonliving components of the natural environment, or technological artifacts, are situated under the word "what." As Derrida (2005, 80) points out, the difference between these two small words already marks/makes a decision concerning "who" will count as morally significant and "what" will and can be excluded as a mere thing. And such a decision is not, it should be emphasized, without ethical presumptions, consequence, and implications.

Chapter 2, therefore, asks whether and under what circumstances machines might be moral patients—that is, someone to whom "who" applies and who, as a result of this, has the kind of moral standing that requires an appropriate response. The conceptual precedent for this reconsideration of the moral status of the machine will be its Cartesian other—the animal. Because the animal and machine traditionally share a common form of alterity—one that had initially excluded both from moral consideration—it would seem that innovations in animal rights philosophy

would provide a suitable model for extending a similar kind of moral respect to the machinic other. This, however, is not the case. Animal rights philosophy, it turns out, is just as dismissive of the machine as previous forms of moral thinking were of the animal. This segregation will not only necessitate a critical reevaluation of the project of animal rights philosophy but will also require that the machine question be approached in a manner that is entirely otherwise.

The third and final chapter responds to the critical complications and difficulties that come to be encountered in chapters 1 and 2. Although it is situated, in terms of its structural placement within the text, as a kind of response to the conflicts that develop in the process of considering moral agency on the one side and moral patiency on the other, this third part of the investigation does not aim to balance the competing perspectives, nor does it endeavor to synthesize or sublate their dialectical tension in a kind of Hegelian resolution. Rather, it approaches things otherwise and seeks to articulate a thinking of ethics that operates beyond and in excess of the conceptual boundaries defined by the terms "agent" and "patient." In this way, then, the third chapter constitutes a *deconstruction* of the agent–patient conceptual opposition that already structures, delimits, and regulates the entire field of operations.

I realize, however, that employing the term "deconstruction" in this particular context is doubly problematic. For those familiar with the continental tradition in philosophy, deconstruction, which is typically associated with the work of Jacques Derrida and which gained considerable traction in departments of English and comparative literature in the United States during the last decades of the twentieth century, is not something that is typically associated with efforts in artificial intelligence, cognitive science, computer science, information technology, and robotics. Don Ihde (2000, 59), in particular, has been critical of what he perceives as "the near absence of conference papers, publications, and even of faculty and graduate student interest amongst continental philosophers concerning what is today often called *technoscience*." Derrida, however, is something of an exception to this. He was, in fact, interested in both sides of the animal-machine. At least since the appearance of the posthumously published *The Animal That Therefore I Am*, there is no question regarding Derrida's interest in the question "of the living and of the living animal" (Derrida 2008, 35). "For me," Derrida (2008, 34) explicitly points out, "that

will always have been the most important and decisive question. I have addressed it a thousand times, either directly or obliquely, by means of readings of all the philosophers I have taken an interest in." At the same time, the so-called father of deconstruction (Coman 2004) was just as interested in and concerned with machines, especially writing machines and the machinery of writing. Beginning with, at least, *Of Grammatology* and extending through the later essays and interviews collected in *Paper Machine*, Derrida was clearly interested in and even obsessed with machines, especially the computer, even if, as he admitted "I know how to make it work (more or less) but I don't know *how* it works" (Derrida 2005, 23).

For those who lean in the direction of the Anglo-American or analytic tradition, however, the term "deconstruction" is enough to put them off their lunch, to use a rather distinct and recognizably Anglophone idiom. Deconstruction is something that is neither recognized as a legitimate philosophical method nor typically respected by mainstream analytic thinkers. As evidence of this, one need look no further than the now famous open letter published on May 9, 1992, in the *Times* of London, signed by a number of well-known and notable analytic philosophers, and offered in reply to Cambridge University's plan to present Derrida with an honorary degree in philosophy. "In the eyes of philosophers," the letter reads, "and certainly among those working in leading departments of philosophy throughout the world, M. Derrida's work does not meet accepted standards of clarity and rigour" (Smith et al. 1992).

Because of this, we should be clear as to what deconstruction entails and how it will be deployed in the context of the machine question. First, the word "deconstruction," to begin with a negative definition, does not mean to take apart, to un-construct, or to disassemble. Despite a popular misconception that has become something of an institutional (mal)practice, it is not a form of destructive analysis, a kind of intellectual demolition, or a process of reverse engineering. As Derrida (1988, 147) described it, "the de- of *de*construction signifies not the demolition of what is constructing itself, but rather what remains to be thought beyond the constructionist or destructionist schema." For this reason, deconstruction is something entirely other than what is understood and delimited by the conceptual opposition situated between, for example, construction and destruction.

Second, to put it schematically, deconstruction comprises a kind of general strategy by which to intervene in this and all other conceptual oppositions that have and continue to organize and regulate systems of knowledge. Toward this end, it involves, as Derrida described it, a double gesture of *inversion* and conceptual *displacement*.

> We must proceed using a double gesture, according to a unity that is both systematic and in and of itself divided, according to a double writing, that is, a writing that is in and of itself multiple, what I called, in "The Double Session," a *double science*. On the one hand, we must traverse a phase of *overturning*. To do justice to this necessity is to recognize that in a classical philosophical opposition we are not dealing with the peaceful coexistence of a *vis-à-vis*, but rather with a violent hierarchy. One of the two terms governs the other (axiologically, logically, etc.), or has the upper hand. To deconstruct the opposition, first of all, is to overturn the hierarchy at a given moment. . . . That being said—and on the other hand—to remain in this phase is still to operate on the terrain of and from the deconstructed system. By means of this double, and precisely stratified, dislodged and dislodging, writing, we must also mark the interval between inversion, which brings low what was high, and the irruptive emergence of a new "concept," a concept that can no longer be, and never could be, included in the previous regime. (Derrida 1981, 41–43)

The third chapter engages in this kind of double gesture or double science. It begins by siding with the traditionally disadvantaged term over and against the one that has typically been privileged in the discourse of the status quo. That is, it initially and strategically sides with and advocates patiency in advance and in opposition to agency, and it does so by demonstrating how "agency" is not some ontologically determined property belonging to an individual entity but is always and already a socially constructed subject position that is "(presup)posited" (Žižek 2008a, 209) and dependent upon an assignment that is instituted, supported, and regulated by others. This conceptual inversion, although shaking things up, is not in and of itself sufficient. It is and remains a mere revolutionary gesture. In simply overturning the standard hierarchy and giving emphasis to the other term, this effort would remain within the conceptual field defined and delimited by the agent–patient dialectic and would continue to play by its rules and according to its regulations. What is needed, therefore, is an additional move, specifically "the irruptive emergence of a new concept" that was not and cannot be comprehended by the previous system. This will, in particular, take the form of another thinking of *patiency* that is not programmed and predetermined as something derived from or the mere

counterpart of agency. It will have been a kind of primordial patiency, or what could be called, following a Derridian practice, an *arche-patient* that is and remains in excess of the agent–patient conceptual opposition.

Questionable Results

This effort, like many critical ventures, produces what are arguably questionable results. This is precisely, as Derrida (1988, 141) was well aware, "what gets on everyone's nerves." As Neil Postman (1993, 181) aptly characterizes the usual expectation, "anyone who practices the art of cultural criticism must endure being asked, What is the solution to the problems you describe?" This criticism of criticism, although entirely understandable and seemingly informed by good "common sense," is guided by a rather limited understanding of the role, function, and objective of *critique*, one that, it should be pointed out, is organized according to and patronizes the same kind of instrumentalist logic that has been applied to technology. There is, however, a more precise and nuanced definition of the term that is rooted in the traditions and practices of critical philosophy. As Barbara Johnson (1981, xv) characterizes it, a critique is not simply an examination of a particular system's flaw and imperfections that is designed to make things better. Instead "it is an analysis that focuses on the grounds of that system's possibility. The critique reads backwards from what seems natural, obvious, self-evident, or universal, in order to show that these things have their history, their reasons for being the way they are, their effects on what follows from them, and that the starting point is not a given but a construct, usually blind to itself" (ibid.). Understood in this way, critique is not an effort that simply aims to discern problems in order to fix them or to ask questions in order to provide answers. There is, of course, nothing inherently wrong with such a practice. Strictly speaking, however, criticism involves more. It consists in an examination that seeks to identify and to expose a particular system's fundamental operations and conditions of possibility, demonstrating how what initially appears to be beyond question and entirely obvious does, in fact, possess a complex history that not only influences what proceeds from it but is itself often not recognized as such.

This effort is entirely consistent with what is called *philosophy*, but we should again be clear as to what this term denotes. According to one way

of thinking, philosophy comes into play and is useful precisely when and at the point that the empirical sciences run aground or bump up against their own limits. As Derek Partridge and Yorick Wilks (1990, ix) write in *The Foundations of Artificial Intelligence*, "philosophy is a subject that comes running whenever foundational or methodological issues arise." One crucial issue for deciding questions of moral responsibility, for example, has been and continues to be *consciousness*. This is because moral agency in particular is typically defined and delimited by a thinking, conscious subject. What comprises consciousness, however, not only is contentious, but detecting its actual presence or absence in another entity by using empirical or objective modes of measurement remains frustratingly indeterminate and ambiguous. "Precisely because we cannot resolve issues of consciousness entirely through objective measurement and analysis (science), a critical role exists for philosophy" (Kurzweil 2005, 380). Understood in this way, philosophy is conceptualized as a supplementary effort that becomes inserted into the mix to address and patch up something that empirical science is unable to answer.

This is, however, a limited and arguably nonphilosophical understanding of the role and function of philosophy, one that already assumes, among other things, that the objective of any and all inquiry is to supply answers to problems. This is, however, not necessarily accurate or appropriate. "There are," Slavoj Žižek (2006b, 137) argues, "not only true or false solutions, there are also false questions. The task of philosophy 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." This effort at reflective self-knowledge is, it should be remembered, precisely what Immanuel Kant, the progenitor of critical philosophy, advances in the *Critique of Pure Reason*, where he deliberately avoids responding to the available questions that comprise debate in metaphysics in order to evaluate whether and to what extent the questions themselves have any firm basis or foundation: "I do not mean by this," Kant (1965, Axii) writes in the preface to the first edition, "a critique of books and systems, but of the faculty of reason in general in respect of all knowledge after which it may strive independently of all experience. It will therefore decide as to the possibility or impossibility of metaphysics in general, and determine its sources, its extent, and its limits." Likewise, Daniel Dennett, who occupies what is often considered to be the opposite

end of the philosophical spectrum from the likes of Žižek and Kant, proposes something similar. "I am a philosopher, 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" (Dennett 1996, vii).

For Kant, Dennett, Žižek, and many others, the task of philosophy is not to supplement the empirical sciences by supplying answers to questions that remain difficult to answer but to examine critically the available questions in order to evaluate whether we are even asking about the right things to begin with. The objective of *The Machine Question*, therefore, will not be to supply definitive answers to the questions of, for example, machine moral agency or machine moral patiency. Instead it will investigate to what extent the way these "problems" are perceived and articulated might already constitute a significant problem and difficulty. To speak both theoretically and metaphorically by way of an image concerning vision ("theory" being a word derived from an ancient Greek verb, *θεωρέω*, meaning "to look at, view, or behold"), it can be said that a question functions like the frame of a camera. On the one hand, the imposition of a frame makes it possible to see and investigate certain things by locating them within the space of our field of vision. In other words, questions arrange a set of possibilities by enabling things to come into view and to be investigated as such. At the same time, and on the other hand, a frame also and necessarily excludes many other things—things that we may not even know are excluded insofar as they already fall outside the edge of what is able to be seen. In this way, a frame also marginalizes others, leaving them on the exterior and beyond recognition. The point, of course, is not simply to do without frames. There is and always will be a framing device of some sort. The point rather is to develop a mode of questioning that recognizes that all questions, no matter how well formulated and carefully deployed, make exclusive decisions about what is to be included and what gets left out of consideration. The best we can do, what we have to and should do, is continually submit questioning to questioning, asking not only what is given privileged status by a particular question and what necessarily remains excluded but also how a particular mode of inquiry already makes, and cannot avoid making, such decisions; what

assumptions and underlying values this decision patronizes; and what consequences—ontological, epistemological, and moral—follow from it. If the machine question is to be successful as a philosophical inquiry, it will need to ask: What is sighted by the frame that this particular effort imposes? What remains outside the scope of this investigation? And what interests and investments do these particular decisions serve?

In providing this explanation, I do not wish to impugn or otherwise dismiss the more practical efforts to resolve questions of responsibility with and by autonomous or semiautonomous machines and/or robots. These questions (and their possible responses) are certainly an important matter for AI researchers, robotics engineers, computer scientists, lawyers, governments, and so on. What I do intend to point out, however, is that these practical endeavors often proceed and are pursued without a full understanding and appreciation of the legacy, logic, and consequences of the concepts they already mobilize and employ. The critical project, therefore, is an important preliminary or prolegomena to these kinds of subsequent investigations, and it is supplied in order to assist those engaged in these practical efforts to understand the conceptual framework and foundation that already structures and regulates the conflicts and debates they endeavor to address. To proceed without engaging in such a critical preliminary is, as recognized by Kant, not only to grope blindly after often ill-conceived solutions to possibly misdiagnosed ailments but to risk reproducing in a supposedly new and original solution the very problems that one hoped to repair in the first place.

3 Thinking Otherwise

3.1 Introduction

Moral philosophy has typically, in one way or another, made exclusive decisions about who is and who is not a legitimate moral agent and/or patient. We have, in effect, sought to determine the line dividing who or what is considered a member of the community of moral subjects from who or what remains outside. And we have done so by, as Thomas Birch (1993, 315) explains, assuming "that we can and ought to find, formulate, establish, institute in our practices, a criterion for (a proof schema of) membership in the class of beings that are moral *consideranda*." It is, for example, no longer news that many if not most of the Western traditions of ethics have been and, in many cases, continue to be exclusively anthropocentric. At the center of mainstream Western ethical theories—irrespective of the different varieties and styles that have appeared under names like virtue ethics, consequentialism, deontologism, or care ethics—has been a common assumption and virtually unquestioned validation of the ἄνθρωπος (*anthropos*)—the ἄνθρωπος who bears a responsibility only to other beings who are like him- or herself, that is, those who are also and already members of the community of ἄνθρωποι. This operation, from one perspective, is entirely understandable and even justifiable insofar as ethical theory is not some transcendent Platonic form that falls fully realized from the heavens but is rather the product of a particular group of individuals, made at a specific time, and initiated in order to protect a particular set of interests. At the same time, however, these decisions have had devastating consequences for others. In other words, any and all attempts to define and determine the proper limit of *consideranda* inevitably proceed by excluding others from participation. "When it comes to

moral considerability," Birch (1993, 315) explains, "there *are*, and *ought* to be, insiders and outsiders, citizens and non-citizens (for example, slaves, barbarians, and women), 'members of the club' of *consideranda* versus the rest."

Ethics, therefore, has been and remains an exclusive undertaking. This exclusivity is fundamental, structural, and systemic. It is not accidental, contingent, or prejudicial in the usual sense of those words. And it is for this reason that little or nothing actually changes as moral theory and practices have developed and matured over time. Even when membership in the club of *consideranda* has, slowly and not without considerable resistance and struggle, been extended to some of these previously excluded others, there have remained other, apparently more fundamental and necessary exclusions. Or to put it another way, every new seemingly progressive inclusion has been made at the expense of others, who are necessarily excluded in the process. Animal rights philosophy, for instance, not only challenged the anthropocentric tradition in ethics but redefined the club of *consideranda* by taking a distinctly animo-centric approach where the qualifying criteria for inclusion in the community of moral subjects was not determined by some list of indeterminate humanlike capabilities—consciousness, rationality, free will, and so on—but the capacity to suffer, or "the ability to not be able," as Derrida (2008, 28) characterizes it.

This effort, despite its important innovations, still excludes others, most notably those nonmammalian animals situated on the lower rungs of the evolutionary ladder; other living organisms like plants and microbes; nonliving natural objects including soils, rocks, and the natural environment taken as a whole; and all forms of nonnatural artifacts, technologies, and machines. And even when these excluded others—these other kinds of others—are finally admitted into the club by other "more inclusive" lists of qualifying criteria that have been proposed by other moral theories, like environmental ethics, machine ethics, or information ethics, exclusions remain. There is, it seems, always someone or something that is and must be *other*. The sequence appears to be infinite, or what Hegel (1987, 137) termed "a bad or negative infinity": "Something becomes an other; this other is itself something; therefore it likewise becomes an other, and so on *ad infinitum*" (ibid.). Ethics, therefore, appears to be unable to do without its others—not only the others who it eventually comes to recognize as Other but also those other others who remain excluded, exterior, and

marginalized. In the final analysis, ethics has been and continues to operate on the basis of a *fraternal logic*—one that defines and defends its membership by always and necessarily excluding others from participation in its exclusive and gated community.

Exclusion is certainly a problem. But inclusion, as its mere flip side and dialectical other, appears to be no less problematic. Despite the recent political and intellectual cachet that has accrued to the word, "inclusion" is not without significant ethical complications and consequences. "The inclusion of the other," as Jürgen Habermas (1998) calls it, whether another human being, animals, the environment, machines, or something else entirely, always and inevitably runs up against the same methodological difficulties, namely the reduction of difference to the same. In order to extend the boundaries of moral agency and/or patiency to traditionally marginalized others, philosophers have argued for progressively more inclusive definitions of what qualifies someone or something for ethical consideration. "The question of considerability has been cast," as Birch (1993, 314) explains by way of Kenneth Goodpaster, "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." The anthropocentric theories, for example, situate the human at the center of ethics and admit into consideration anyone who is able to meet the basic criteria of what has been decided to constitute the human being—even if, it should be recalled, this criterion has itself been something that is arguably capricious and not entirely consistent. Animal rights philosophy focuses attention on the animal and extends consideration to any organism that meets its defining criterion of "can they suffer?" The biocentric efforts of some forms of environmental ethics go one step further in the process, defining life as the common denominator and admitting into consideration anything and everything that can be said to be alive. And ontocentrism completes the expansion of moral consideration by incorporating anything that actually exists, had existed, or potentially exists, and in this way, as Floridi (1999) claims, provides the most universal and totalizing form of an all-inclusive ethics.

All these innovations, despite their differences in focus and scope, employ a similar maneuver. That is, they redefine the center of moral consideration in order to describe progressively larger and more inclusive circles that are able to encompass a wider range of possible participants.

Although there are and will continue to be considerable disagreements about who or what should define the center and who or what is or is not included, this debate is not the problem. The problem rests in the strategy itself. In taking this particular approach, these different ethical theories endeavor to identify what is essentially the same in a phenomenal diversity of individuals. Consequently, they include others by effectively stripping away and reducing differences. This approach, although having the appearance of being increasingly more inclusive, "is rather clearly a function of imperial power mongering," as Birch (1993, 315) describes it. For it immediately effaces the unique alterity of others and turns them into more of the same, instituting what Slavoj Žižek (1997, 161) calls the structure of the Möbius band: "At the very heart of Otherness, we encounter the other side of the Same." In making this argument, however, it should be noted that the criticism has itself employed what it criticizes. (Or to put it another way, the articulation of what is the matter is itself already and unavoidably involved with the material of its articulation.) In focusing attention on what is essentially the same in these various forms of moral centrism, the analysis does exactly what it charges—it identifies a common feature that underlies apparent diversity and effectively reduces a multiplicity of differences to what is the same. Pointing this out, however, does not invalidate the conclusion but demonstrates, not only in what is said but also in what is done, the questionable operations that are already involved in any attempt at articulating inclusion.

Exclusion is a problem because it calls attention to and fixates on what is different despite what might be similar. Inclusion is a problem, because it emphasizes similarities at the expense of differences. Consequently, the one is the inverse of the other. They are, as Michael Heim (1998, 42) calls them, "binary brothers," or, to put it in colloquial terms, two sides of one coin. As long as moral debate and innovation remain involved with and structured by these two possibilities, little or nothing will change. Exclusion will continue to be identified and challenged, as it has been in the discourses of moral personhood, animal rights, bioethics, and information ethics, by calls for greater inclusiveness and ethical theories that are able to accommodate these previously excluded others. At the same time, efforts to articulate inclusion will be challenged, as they have been in critical responses to these projects, as "imperial power mongering" (Birch 1993, 315) and for the reduction of difference that they had sought to respect and accommodate in the first place.

What is needed, therefore, is a third alternative that does not simply oppose exclusion by inclusion or vice versa. What is needed is an approach that is situated and oriented otherwise. In thinking otherwise, we will not be interested in taking sides or playing by the existing rules of the game. Instead we will be concerned with challenging, criticizing, and even changing the terms and conditions by which this debate has been organized, articulated, and configured. Precedent for this kind of alternative transaction can already be found in both the continental and analytic traditions. It is, for example, what poststructuralists like Jacques Derrida propose with the term "deconstruction," and what Thomas Kuhn endeavors to articulate in his paradigm changing work *The Structure of Scientific Revolutions*. But the practice is much older. It is, for instance, evident in Immanuel Kant's "Copernican revolution," which sought to resolve fundamental questions in modern philosophy not by lending support to or endeavoring to prove one or the other side in the rationalist versus empiricist debate but by rewriting rules of the game (Kant 1965, Bxvi).

But thinking otherwise is even older than this innovation in modern European thought, having its proper origin in the inaugural gesture attributed to the first philosopher, Socrates. It is, as John Sallis points out, in Plato's *Phaedo*, a dialogue that narrates among other things the final hours of Socrates's life, that the aged philosopher remembers where it all began. "In the face of death Socrates recalls how he became what he is: how he began by following the ways of his predecessors, the ways handed down through the operation of a certain tradition, the way of that kind of wisdom called *περί φύσες ίστορία*; how this alleged wisdom repeatedly left him adrift; and how, finally taking to the oars, he set out on a second voyage by having recourse to *λόγοι*" (Sallis 1987, 1). As long as Socrates followed the tradition he had inherited from his predecessors—asking the questions they had already determined to be important, following the methods they had prescribed as being the most effective, and evaluating the kind of evidence they would recognize as appropriate—he failed. Rather than continue on this arguably fruitless path, he decides to change course by proceeding and thinking otherwise.

3.2 Decentering the Subject

Although not a Platonist by any means or even a philosopher, the roboticist Rodney Brooks supplies one effective method for pursuing the

alternative of thinking otherwise. In *Flesh and Machines: How Robots Will Change Us*, Brooks describes, by way of an autobiographical gesture that is not unlike the one Plato has Socrates deploy, the "research heuristic" that was largely responsible for much of his success: "During my earlier years as a postdoc at MIT, and as a junior faculty member at Stanford, I had developed a heuristic in carrying out research. I would look at how everyone else was tackling a certain problem and find the core central thing that they had all agreed on so much that they never even talked about it. Then I would negate the central implicit belief and see where it led. This often turned out to be quite useful" (Brooks 2002, 37).

Following this procedure, we can say that one of the common and often unacknowledged features of the different formulations of both moral agency and patiency is the assumption that moral considerability is something that can and should be decided on the basis of individual qualities. This core assumption is clearly operational in, for example, the question of moral personhood, where the principal objective is to identify or articulate "the person-making qualities" in a way that is nonarbitrary and nonprejudicial; demonstrate how something, say, an animal or a machine, does in fact provide evidence of possessing that particular set of qualities; and establish guidelines that specify how such persons should be treated by others in the group. Even though there remain considerable disagreements about the exact qualities or criteria that should apply, what is not debated is the fact that an individual, in order to be considered a legitimate moral person, would need to achieve and demonstrate possession of the necessary and sufficient conditions for inclusion in the club. Instead of continuing in this fashion, arguing that some other individuals also clear the bar or making a case to revise the criteria of inclusion, we can proceed otherwise. Specifically, we can challenge or "negate," which is Brook's term, the basic assumption concerning the privileged place of the individual moral subject, arguably a product of Cartesian philosophy and the enlightenment's obsession with the self, with a decentered and distributed understanding of moral subjectivity. We can, in effect, agree that the center always and already cannot hold and that "things fall apart" (Yeats 1922, 289; Achebe 1994).

One such alternative can be found in what F. Allan Hanson (2009, 91) calls "extended agency theory," which is itself a kind of extension of actor-network approaches. According to Hanson, who takes what appears

to be a practical and entirely pragmatic view of things, machine responsibility is still undecided and, for that reason, one should be careful not to go too far in speculating about the issue: "Possible future development of automated systems and new ways of thinking about responsibility will spawn plausible arguments for the moral responsibility of non-human agents. For the present, however, questions about the mental qualities of robots and computers make it unwise to go this far" (ibid., 94). Instead, Hanson, following the work of Peter-Paul Verbeek (2009), suggests that this problem may be resolved by considering various theories of "joint responsibility," where "moral agency is distributed over both human and technological artifacts" (Hanson 2009, 94). This is an elaboration of the "many hands" concept that had been proposed by Helen Nissenbaum (1996) to describe the distributed nature of accountability in computerized society.

In this way, Hanson's "extended agency" introduces a kind of "cyborg moral subject" where responsibility resides not in a predefined ethical individual but in a network of relations situated between human individuals and others, including machines. For Hanson, this distributed moral responsibility moves away from the anthropocentric individualism of enlightenment thought, which divides the world into self and other, and introduces an ethics that is more in line with recent innovations in ecological thinking:

> When the subject is perceived more as a verb than a noun—a way of combining different entities in different ways to engage in various activities—the distinction between Self and Other loses both clarity and significance. When human individuals realize that they do not act alone but together with other people and things in extended agencies, they are more likely to appreciate the mutual dependency of all the participants for their common well-being. The notion of joint responsibility associated with this frame of mind is more conducive than moral individualism to constructive engagement with other people, with technology, and with the environment in general. (Hanson 2009, 98)

A similar proposal is provided in David F. Channell's *The Vital Machine*. After a rather involved investigation of the collapse of the conceptual differences customarily situated between technology and organisms, Channell ends his analysis with a brief consideration of "ethics in the age of the vital machine." "No longer," Channell (1991, 146) argues, beginning with a characterization that deliberately negates the standard approach, "can the focus of the theory of ethics be the autonomous individual. No longer

can ethical judgments be based on a simple distinction between the intrinsic value of human beings and the instrumental value of technological creations. The focus of the ethics of the vital machine must be decentered." This decentering, however, does not go for all machines in all circumstances. Instead, it is contextualized in such a way as to be responsive to and responsible for differences in particular situations: "In some cases, with the use of traditional tools, the interactions may be very simple and the 'center' of ethics will be more on the side of the human, but in other cases, with the use of intelligent computers, the interactions may be quite complex and the 'center' of ethics will be more or less equally divided between the human and the machine" (ibid.). To respond to these apparent shifts in the 'center' of moral consideration, Channell proposes "a decentered ethical framework that reflects a bionic world view" (ibid., 152), what he calls "a bionic ethic" (ibid., 151).

This idea is derived from a reworking of Aldo Leopold's (1966) "land ethic." "While the land ethic of Leopold focuses on the organic, and in fact is usually interpreted as being in opposition to technology, it does provide a model for including both the organic and the mechanical into the expanding boundaries of a new ethic. In point of fact, Leopold often explained the interdependence of the biotic elements of nature in terms of engine parts or wheels and cogs" (Channell 1991, 153). Although often distinguished from technological concerns, Channell finds Leopold's land ethic to provide articulation of a moral thinking that can respect and take responsibility for nonliving objects, not only soils, waters, and rocks but also computers and other technological artifacts. For Channell, connecting the dots between these different concerns is not only a matter of metaphorical comparison—that is, the fact that nature has often been described and characterized in explicit mechanical terms—but grounded in established moral and legal precedent, that is, in the fact that "inanimate objects such as trusts, corporations, banks, and ships have long been seen by the courts as possessing rights"; the fact that some "writers have suggested that landmark buildings should be treated in a way similar to endangered species"; and the fact that "objects of artistic creation . . . have an intrinsic right to exist and be treated with respect" (ibid.).

In taking this approach, however, Channell is not arguing that inanimate objects and artifacts, like machines, should be considered the same as human beings, animals, or other living organisms. Instead, following

Leopold, he advocates a holistic ecological perspective, something that is called, borrowing a term from Richard Brautigan, "a cybernetic ecology." "The idea of a cybernetic ecology," Channell argues, "does not imply that machines should be given equal standing with humans or with animals, plants, rivers, or mountains. Even within nature, there is a hierarchy of living things, with some species dominant over others. A fundamental element of any ecological system is the 'food chain.' Throughout the environment the continued survival of one species is dependent on its being able to eat (or in more general terms transfer energy from) another part of the environment" (ibid., 154). The main issue, therefore, is figuring out where the various technological artifacts fit in the "food chain" of this "cybernetic ecology." Although this is, for now at least, still a largely undecided issue, what Channell proposes is a much more holistic approach and understanding of the moral landscape. For him, the issue is not simply who is and who is not part of an exclusive club, but rather how the different elements of the ecology fit together and support each other in a system that includes not just "deers and pines" but also "computers and electronics" (ibid.). "Within an ecological system, all elements have some intrinsic value but because of the interdependence within the system every element also has some instrumental value for the rest of the system. Each part of the ecology has a certain degree of autonomy, but in the context of the system, each part plays some role in the control of the entire ecology" (ibid.). What Channell advocates, therefore, is a shift in perspective from a myopic Cartesian subject to a holistic ecological orientation in which each element becomes what it is as a product of the position it occupies within the whole and is granted appropriate rights and responsibilities in accordance with the functioning and continued success of the entire system.

This decentered, systems approach to deciding questions of moral considerability sounds promising, but it has problems. First, this new cybernetic holism (and it should be recalled that cybernetics, from the very beginning, was to have been a totalizing science covering both animals and machines), which leverages the land ethic of Leopold, inherits many of the problems typically associated with environmental ethics. Although it challenges and deposes the anthropocentric privilege that human beings had traditionally granted themselves in moral philosophy, it still, as Birch (1993, 315) points out, locates a center of moral concern and organizes

and regulates its system of ethics according to this new moral subject. Despite significantly challenging the anthropocentric perspective, this shift in focus is still and cannot help but be centrist. It simply redistributes what is considered to be the center of the moral universe. So for all its promise to decenter things, Channell's bionic ethic is just one more in a long line of competing and more inclusive forms of centrisms. Like Floridi's IE, it is clearly more universal and more inclusive, but it is, on this account, just more of the same.

Second, there is, in all of this, a problem with subjectivity. This comes out in the final paragraph of Channell's text, where he ends on an arguably optimistic if not utopian note: "In a cybernetic ecology both technology and organic life must be intelligently conserved. The entire system might be worse off without the peregrine falcon or the snail darter, but it also might be worse off without telecommunications and much of medical technology. On the other hand we might not want to conserve nuclear weapons or dioxin, but we might also be better off if the AIDS virus became extinct. In the end, we will build a new Jerusalem only if we can find a harmony between organic life and technology" (Channell 1991, 154). What remains unanswered in this optimistic assessment is: Who or what is the subject of this passage? Who or what is marked with the pronoun "we?" Who or what speaks in this conclusive statement? If the first-person plural refers to human beings, if it addresses itself to those individual humans who read the text and share a certain language, community, and tradition, then this statement, for all its promise, seems to sneak anthropocentrism in the back door. In this way, the cybernetic ecological perspective would become just another way of articulating, preserving, and protecting what are, in the final analysis, human interests and assets. And this conclusion seems to be supported by the examples Channell provides, insofar as the AIDS virus is something that adversely affects the immune system and integrity of the human species.

It is also possible, however, that this "we" refers not to human beings but to the vital machine and the entire cybernetic ecology. But then the issue must be who or what gives Channell, presumably a human being, the right to speak on behalf of this larger community. By what right does this individual, or anyone else, for that matter, write and speak on behalf of all the members of this community—human, animal, machine,

or otherwise? Who or what grants this authority to speak, in this particular idiom, on behalf of this larger whole? This is of course the problem with any form of religious discourse, in which a particular human individual, like a prophet, or group of human beings, like a church, speaks on behalf of the divine, articulating what god wants, needs, or desires. Doing so is clearly expedient, but it is also a decisive imposition of power—what Birch (1993, 315) calls "imperial power mongering." Consequently, if Channell's "we" is human, it is not enough. If, however, his use of this term refers to the vital machine or the whole cybernetic ecology, it is perhaps too much. In pointing this out, my goal is not to fault Channell for getting it wrong but to point out how trying to get it right is already constrained and limited by the very system against which one struggles. Proposing an alternative, therefore, is neither simple, complete, nor beyond additional critical reflection.

An alternative thinking of decentered ethics is proposed by Johanna Zylinska (2009, 163), who advocates "a Deleuzian-influenced notion of 'distributive and composite agency.'" This form of "distributive agency" is proposed in direct response to and as an alternative for contemporary metaethical innovations that, although critical of the anthropocentric tradition, unfortunately do not go far enough. Zylinska, for instance, argues that the apparently groundbreaking work of animal rights philosophers, like Peter Singer, succeeds in "radicalizing humanist ethics by shifting the boundaries of who counts as a 'person'" while it "still preserves the structural principles of this ethics, with an individual person serving as the cornerstone" (ibid., 14). According to Zylinska, therefore, Singer merely remixes and modifies traditional anthropocentric ethics. He questions who or what gets to participate but ultimately preserves the basic structure and essential rules of the humanist game. In contrast, a concept of "distributive agency" recognizes and affirms the fact that an "individual person," however that comes to be defined, is always situated in and already operating through complex networks of interacting relations.

"Human agency," Zylinska argues with reference to the cyborg performance art of Stelarc, "does not disappear altogether from this zone of creative and contingent evolution, but it is distributed throughout a system of forces, institutions, bodies, and nodal points. This acknowledgement of agential distribution—a paradox that requires a temporarily rational and self-present self which is to undertake this realization—allows for an

enactment of a more hospitable relationship to technology than the paranoid fear of the alien other" (ibid., 172). In this way, then, "distributive and composite agency" or what Zylinska also calls an "agency of assemblages" (ibid., 163) goes beyond the "many hands" thesis of Nissenbaun, Hanson's "extended agency theory," or Channell's "bionic ethic." Whereas these other theorists advocate the decentering of agency within a network of actors, Zylinska uses this distribution as a way to develop a distinctly Levinasian-inspired form of hospitality for others—one that can remain open to a completely different and alien other. In other words, where other forms of a decentered ethics inevitably focus attention on some other center—relying on the very structural gesture and approach that they had wanted to contest in the first place—Zylinska proposes a radically decentering in which nothing is a center of moral concern but everything can potentially be subject to ethics. What makes Zylinska's decentering work is its attention to the exorbitant other and other forms of otherness. This alternative way of looking at things ultimately concerns and will need to be referred to a reformulation of the question of moral patiency. In fact, this alternative approach seems to make patiency the privileged term rather than a derived aspect of some predefined notion of moral agency. Someone or something becomes a moral agent only after first being admitted into the fraternity of moral subjects—only after and on the condition that some other is recognized as Other.

3.3 The Ethics of Social Construction

A decentered approach recognizes the way moral responsibility is often not constituted by an individual subject but instead comes to be distributed across a network of interrelated and interacting participants. But this is not the only way to proceed. Other alternatives focus not on a decentering of the moral subject by tracing its distribution within the social fabric of a network but consider how the moral subject, whether conceptualized as an agent, patient, or both, has been socially constructed, regulated, and assigned. One such alternative is advanced by Bernd Carsten Stahl under the concept of "quasi-responsibility." In the article "Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency," Stahl effectively skirts the question of agency and personhood by reformulating the entire approach: "Instead of