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

For Ann on Mother's Day, 2011

Preface

At one time I had considered titling this book *A Vindication of the Rights of Machines*, for two reasons. First, such a designation makes reference to and follows in the tradition of "vindication discourses," if one might be permitted such a phrase, that begins with Mary Wollstonecraft's *A Vindication of the Rights of Men* (1790) followed 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*, also published in the year 1792. Following suit, this book inquires about and advances the question concerning the possibility of extending rights and responsibilities to machines, thereby comprising what would be the next iteration in this lineage of discourses addressing the rights of previously excluded others.

The second reason was that I had previously employed the title "The Machine Question" in another book, *Thinking Otherwise: Philosophy, Communication, Technology* (Gunkel 2007), as the heading to that text's final chapter. And it is always good strategy to avoid this kind of nominal repetition even if, as is the case, this undertaking is something of a sequel, extension, and elaboration of that previous effort. To complicate matters and to "return the favor," this book ends with a chapter called, quite deliberately, "thinking otherwise," which has the effect of transforming what had come before into something that now can be read as a kind of sequel. So using the "vindication" moniker would have helped minimize the effect of this mirror play.

But I eventually decided against this title, again for two reasons. First, "vindication discourses" are a particular kind of writing, similar to a manifesto. The opening lines of Taylor's text indicate the kind of tone and rhetoric that is to be expected of such an undertaking: "It appears at first sight somewhat singular, that a moral truth of the highest importance, and most illustrious evidence, should have been utterly unknown to the ancients, and not yet fully perceived, and universally acknowledged, even in such an enlightened age as the present. The truth I allude to is, *the equality of all things, with respect to their intrinsic and real dignity and worth*" (Taylor 1966, 9). There is nothing in the following that approaches this kind of direct and bold declaration of self-evident and indubitable truths. For even that approach needs to be and will be submitted to critical questioning. Consequently, the moniker *A Vindication of the Rights of Machines*, as useful as it first seems, would have been a much more accurate description of the final chapter to *Thinking Otherwise*, which dissimulates this kind of rhetoric in an attempt to make a case for the advancement of the rights of machines in opposition to the anthropocentric tradition in moral philosophy.

Second, the title *The Machine Question* not only makes reference to and leverages the legacy of another moral innovation—one that has been situated under the phrase "the animal question"—but emphasizes the role and function of *questioning*. Questioning is a particularly philosophical enterprise. Socrates, as Plato describes in the *Apology*, does not get himself into trouble by making claims and proclaiming truths. He simply investigates the knowledge of others by asking questions (Plato 1990, 23a). Martin Heidegger, who occupies a privileged position on the continental side of the discipline, begins his seminal *Being and Time* (1927) not by proposing to answer "the question of being" with some definitive solution, but by attending to and renewing interest in the question: "Haben wir heute eine Antwort auf die Frage nach dem, was wir mit dem Wort 'seiend' eigentlich meinen? Keineswegs. Und so gilt es denn, *die Frage nach dem Sinn von Sein* erneut zu stellen [Do we in our time have an answer to the question of what we really mean by the word 'being?' Not at all. So it is fitting that we should raise anew *the question of the meaning of Being*]" (Heidegger 1962, 1). And on the other side of the philosophical divide, G. E. Moore, whom Tom Regan (1999, xii) called "analytic philosophy's patron saint," takes a similar approach, writing the following in, of all places, the preface to his influential *Principia Ethica* (1903): "It appears to me that in Ethics, as in all other philosophical studies, the difficulties and disagreements, of which its history is full, are mainly due to a very simple cause: namely to the attempt to answer questions, without first discovering precisely *what* question it is which you desire to answer" (Moore 2005, xvii).

In the end, I decided on the title *The Machine Question*, precisely because what follows draws on, is dedicated to, and belongs to this philosophical lineage. As such, the analysis presented in this book does not endeavor to answer the question concerning the moral status of the machine with either a "yes" or "no." It does not seek to prove once and for all that a machine either can be or cannot be a legitimate moral subject with rights and responsibilities. And it does not endeavor to identify or to articulate moral maxims, codes of conduct, or practical ethical guidelines. Instead it seeks *to ask the question*. It endeavors, as Heidegger would describe it, to learn to attend to the machine question in all its complexity and in the process to achieve the rather modest objective, as Moore describes it, of trying to discover what question or questions we are asking before setting out to try to supply an answer. For this reason, if *The Machine Question* were to have an epigraph, it would be these two opening statements from Heidegger and Moore (two philosophers who could not be more different from each other), concerning the role, function, and importance of questioning.

Acknowledgments

Much of the material included in this book was originally formulated in response to opportunities, provocations, and challenges offered by Richard Johannesen and Clifford Christians. Its structure initially took shape in the process of participating in a conference panel with Heidi Campbell and Lesley Dart. And it all came together under the skillful direction of Philip Laughlin at the MIT Press.

This book, perhaps more so than my others, bears the distinct imprint of those individuals who introduced me to philosophy: Jim Cheney got me thinking about others and other kinds of others; Terry Penner provided access to Plato and helped me appreciate the analytic tradition; John Sallis, my Master's advisor, taught me how to read the major figures of continental thought, especially Kant, Hegel, Heidegger, and Derrida; and David Farrell Krell, my *Doktorvater*, taught me how to write and make it all work . . . but with a distinctly Nietzschean sense of play. Danke sehr!

Two colleagues on the other side of the Atlantic provided ongoing support and insight throughout the project: Paul Taylor, my partner in crime at the *International Journal of Žižek Studies*, has been and continues to be a patient sounding-board for all kinds of things. And Joanna Bryson, who I first met by chance in the laundry room of the Grandeur Apartment building (1055 W. Granville Ave., Chicago, Illinois) in the mid-1980s, has continued to challenge and influence my thinking about computers and robots even if we come at this stuff from very different perspectives.

The final chapter got a major boost from engaging conversations with colleagues in Brazil. These interactions came at just the right time and helped reorient a good deal of that material. I am especially grateful to Ciro Marcondes Filho of Escola de Comunicações e Artes, University of São Paulo, for the invitation to participate in the "10 anos de FiloCom" conference, and to the following scholars who contributed, in one way or another, to the conversation: Marco Toledo Bastos, Cristina Pontes Bonfiglioli, Massimo Di Felice, Maurício Liesen, Danielle Naves de Oliveira, Francisco Rüdiger, Liv Sovik, and Eugênio Trivinho. Obrigado!

The structure and presentation of the text has benefited greatly from the experience of writing a failed grant application and the insightful conversations that that exercise occasioned with David Stone and Andrea Buford of the Northern Illinois University (NIU) Office of Sponsored Projects. I have also had the opportunity to work with two talented research assistants. Jennifer Howard of NIU Media Services produced the trailer for the book, which is available at http://machinequestion.org, and Michael Gracz helped out with research tasks and manuscript preparation. I also acknowledge my colleagues in the Department of Communication at NIU who continue to provide a supportive environment in which to think, work, and write. This is absolutely essential and greatly appreciated.

This book would not have been possible without the continued support and love of my family, my wife Ann Hetzel Gunkel and son Stanisław Gunkel. Youse [*sic*] make everyday a joy, even if things like hockey practice, violin lessons, etc. interrupt the writing. I wouldn't have it any other way. Dzięki serdeczne!

Finally, I would like to express my gratitude to every machine that assisted or participated in the production of this book. Although I have no way of knowing whether you know it or not, I could not have done it without you. 01110100 01101000 01100001 01101110 01101011 01110011

Machinequestion.org

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 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. Theodor 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