Cambridge University Press
978-0-521-11235-2- Machine Ethics
Edited by Michael Anderson and Susan Leigh Anderson
Excerpt
More information

General Introduction

The subject of this book is a new field of research: developing ethics for machines, in contrast to developing ethics for human beings who use machines. The distinction is of practical as well as theoretical importance. Theoretically, *machine ethics* is concerned with giving *machines* ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making. In the second case, in developing ethics for human beings who use machines, the burden of making sure that machines are never employed in an unethical fashion always rests with the *human beings* who interact with them. It is just one more domain of applied *human* ethics that involves fleshing out proper and improper *human* behavior concerning the use of machines. Machines are considered to be just tools used by human beings, requiring ethical guidelines for how they ought and ought not to be used by humans.

Practically, the difference is of particular significance because succeeding in developing ethics for machines enables them to function (more or less) *autonomously*, by which is meant that they can function without human causal intervention after they have been designed for a substantial portion of their behavior. (Think of the difference between an ordinary vacuum cleaner that is guided by a human being who steers it around a room and a Roomba that is permitted to roam around a room on its own as it cleans.) There are many necessary activities that we would like to be able to turn over entirely to autonomously functioning machines, because the jobs that need to be done are either too dangerous or unpleasant for humans to perform, or there is a shortage of humans to perform the jobs, or machines could do a better job performing the tasks than humans. Yet no one would feel comfortable allowing machines to function autonomously without ethical safeguards in place. Humans could not micromanage the behavior of the machines without sacrificing their ability to function autonomously, thus losing the benefit of allowing them to replace humans in performing certain tasks. Ideally, we would like to be able to trust autonomous machines to make correct ethical decisions on their own, and this requires that we create an ethic for machines.

It is not always obvious to laypersons or designers of machines that the behavior of the sort of machines to which we would like to turn over necessary or desired tasks has ethical import. If there is a possibility that a human being could be harmed should the machine behave in a certain manner, then this has to be taken into account. Even something as simple as an automatic cash-dispensing machine attached to a bank raises a number of ethical concerns: It is important to make it extremely difficult for the cash to be given to a person other than the customer from whose account the money is withdrawn; but if this should happen, it is necessary to ensure that there will be a way to minimize the harm done both to the customer and the bank (harm that can affect many persons' lives), while respecting the privacy of the legitimate customer's transactions and making the machine easy for the customer to use.

From just this one example, we can see that it will not be easy to incorporate an ethical dimension into autonomously functioning machines. Yet an automatic cash-dispensing machine is far less complex – in that the various possible actions it could perform can be anticipated in advance, making it relatively simple to build ethical safeguards into its design – than the sort of autonomous machines that are currently being developed by AI researchers. Adding an ethical component to a complex autonomous machine, such as an eldercare robot, involves training a machine to properly weigh a number of ethically significant factors in situations not all of which are likely to be anticipated by their designers.

Consider a demonstration video of a robot currently in production that raises ethical concerns in even the most seemingly innocuous of systems. The system in question is a simple mobile robot with a very limited repertoire of behaviors that amount to setting and giving reminders. A number of questionable ethical practices can be discerned in the demonstration. For instance, after asking the system's charge whether she had taken her medication, the robot asks her to show her empty pillbox. This is followed by a lecture by the robot concerning how important it is for her to take her medication. There is little back story provided, but assuming a competent adult, such paternalistic behavior seems uncalled for and shows little respect for the patient's autonomy.

During this exchange, the patient's responsible relative is seen watching it over the Internet. Although it is not clear whether this surveillance has been agreed to by the person being watched – there is no hint in the video that she knows she is being watched – there is the distinct impression left that her privacy is being violated.

As another example, promises are made by the system that the robot will remind its charge when her favorite show and "the game" are on. Promise making and keeping clearly have ethical ramifications, and it is not clear that the system under consideration has the sophistication to make ethically correct decisions when the duty to keep promises comes into conflict with other possibly more important duties.

Finally, when the system does indeed remind its charge that her favorite television show is starting, it turns out that she has company and tells the robot to go away. The robot responds with "You don't love me anymore," to the delight of the guests, and slinks away. This is problematic behavior because it sets up an expectation in the user that the system cannot fulfill – that it is capable of a loving relationship with its charge. This is a very highly charged ethical ramification, particularly given the vulnerable population for which this technology is being developed.

The bottom line is that, contrary to those who argue that concern about the ethical behavior of autonomous systems is premature, the behavior of even the simplest of such systems such as the one in our example shows that, in fact, such concern is overdue. This view has recently been expressed by Great Britain's Royal Academy of Engineering in the context of domestic autonomous systems: "Smart homes are close to the horizon and could be of significant benefit. However, they are being developed largely without ethical research. This means that there is a danger of bad design, with assumptions about users and their behavior embedded in programming. It is important that ethical issues are not left for programmers to decide – either implicitly or explicitly."

Developing ethics for machines requires research that is interdisciplinary in nature. It must involve a dialogue between ethicists and specialists in artificial intelligence. This presents a challenge in and of itself, because a common language must be forged between two very different fields for such research to progress. Furthermore, there must be an appreciation, on both sides, of the expertise of the other. Ethicists must accept the fact that there can be no vagueness in the programming of a machine, so they must sharpen their knowledge of ethics to a degree that they may not be used to. They are also required to consider real-world applications of their theoretical work. Being forced to do this may very well lead to the additional benefit of advancing the field of ethics. As Daniel Dennett recently stated, "AI makes Philosophy honest."

AI researchers working on machine ethics, on the other hand, must accept that ethics is a long-studied discipline within the field of philosophy that goes far beyond laypersons' intuitions. Ethicists may not agree on every matter, yet they have made much headway in resolving disputes in many areas of life. Agreed upon, all-encompassing ethical principles may still be elusive, but there is much agreement on acceptable behavior in many particular ethical dilemmas, hopefully in the areas where we would like autonomous machines to function. AI researchers need to defer to ethicists in determining when machine behavior raises ethical concerns and in making assumptions concerning acceptable machine behavior. In areas where ethicists disagree about these matters, it would be unwise to develop machines that function autonomously.

The essays in this volume represent the first steps by philosophers and AI researchers toward explaining why it is necessary to add an ethical dimension to machines that function autonomously; what is required in order to add this dimension; philosophical and practical challenges to the machine ethics project; various approaches that could be considered in attempting to add an ethical dimension to machines; work that has been done to date in implementing these approaches; and visions of the future of machine ethics research.

The book is divided into five sections. In the first section, James Moor, Susan Leigh Anderson, and J. Storrs Hall discuss the nature of machine ethics, giving an overview of this new field of research. In the second section, Colin Allen, Wendell Wallach, Iva Smit, and Sherry Turkel argue for the importance of machine ethics. The authors in the third section of the book – Drew McDermott, Steve Torrance, Blay Whitby, John Sullins, Susan Leigh Anderson, Deborah G. Johnson, Luciano Floridi, and David J. Calverley – raise issues concerning the machine ethics agenda that will need to be resolved if research in the field is to progress. In the fourth section, various approaches to capturing the ethics that should be incorporated into machines are considered and, for those who have begun to do so, how they may be implemented. James Gips gives an overview of many of the approaches. The approaches that are considered include: Asimov's Laws, discussed by Roger Clarke and Susan Leigh Anderson; artificial intelligence approaches, represented in the work of Bruce McLaren, Marcello Guarini, Alan K. Mackworth, Selmer Bringsjord et al., Matteo Turilli, Luís Moniz Pereira and Ari Saptawijaya; psychological/sociological approaches, represented in the work of Morteza Dehghani, Ken Forbus, Emmett Tomai, Matthew Klenk, and Peter Danielson; and philosophical approaches, discussed by Christopher Grau, Thomas M. Powers, and Susan Leigh Anderson and Michael Anderson. Finally, in the last section of the book, four visions of the future of machine ethics are given by Helen Seville, Deborah G. Field, J. Storrs Hall, Susan Leigh Anderson, and Eric Dietrich.

Part I

The Nature of Machine Ethics

Introduction

James Moor, in "The Nature, Importance, and Difficulty of Machine Ethics," discusses four possible ways in which values could be ascribed to machines. First, ordinary computers can be considered to be "normative agents" but not necessarily *ethical* ones, because they are designed with a purpose in mind (e.g., to prove theorems or to keep an airplane on course). They are technological agents that perform tasks on our behalf, and we can assess their performance according to how well they perform their tasks. Second, "ethical impact agents" not only perform certain tasks according to the way they were designed, but they also have an ethical impact (ideally a positive one) on the world. For example, robot jockeys that guide camels in races in Qatar have replaced young boys, freeing them from slavery. Neither of the first two senses of ascribing values to machines, Moor notes, involves "putting ethics into a machine," as do the next two.

Third, "implicit ethical agents" are machines that have been programmed in a way that supports ethical behavior, or at least avoids unethical behavior. They are constrained in their behavior by their *designers* who *are following ethical principles*. Examples of such machines include ATMs that are programmed not to cheat the bank or its customers and automatic airplane pilots that are entrusted with the safety of human beings. Moor maintains that good software engineering should include requiring that ethical considerations be incorporated into machines whose behavior affects human lives, so at least this sense of "machine ethics" should be accepted by all as being desirable.

Fourth, "explicit ethical agents" are able to calculate the best action in ethical dilemmas. These machines would be able to "do ethics in a way that, for example, a computer can play chess." They would need to be able to represent the current situation, know which actions are possible in this situation, and be able to assess these actions in terms of some ethical theory, enabling them to calculate the ethically best action, just as a chess-playing program can represent the current board positions, know which moves are legal, and assess these moves in terms of achieving the goal of checkmating the king, enabling it to figure out the best move. Is it possible to create such a machine? Moor agrees with James Gips that "the development of a machine that's an explicit ethical agent seems a fitting subject for a [computing] Grand Challenge."

Most would claim that even if we could create machines that are explicit ethical agents, we would still not have created what Moor calls "full ethical agents," a term used to describe human ethical decision makers. The issue, he says, is whether intentionality, consciousness, and free will – attributes that human ethical agents possess or are at least thought to possess – are essential to genuine ethical decision making. Moor wonders whether it would be sufficient that machines have "as if it does" versions of these qualities. If a machine is able to give correct answers to ethical dilemmas and even give justifications for its answers, it would pass Colin Allen's "Moral Turing Test" (Allen et al.: Prolegomena to any future artificial moral agent. *J. Exp. Theor. Artif. Intell*. 12(3): 251–261) for "understanding" ethics. In any case, we cannot be sure that machines that are created in the future will lack the qualities that we believe now uniquely characterize human ethical agents.

Anticipating the next part of the book, Moore gives three reasons "why it's important to work on machine ethics in the sense of developing explicit ethical agents": (1) because ethics itself is important, which is why, at the very least, we need to think about creating *implicit* ethical machines; (2) because the machines that are being developed will have increasing autonomy, which will eventually force us to make the ethical principles that govern their behavior *explicit* in these machines; and (3) because attempting to program ethics into a machine will give us the opportunity to understand ethics better.

Finally, Moor raises three concerns with the machine ethics project that should be considered in connection with the third part of the book: (1) We have a limited understanding of ethics. (2) We have a limited understanding of how learning takes place. (3) An ethical machine would need to have better "common sense and world knowledge" than computers have now.

Most of what Moor has to say would appear to be noncontroversial. Steve Torrance, however, has argued (in his paper "A Robust View of Machine Ethics," proceedings of the AAAI Fall Symposium on Machine Ethics, 2005), in contrast to Moor's view that the machines created in the future may have the qualities we believe are unique to human ethical agents, that to be a full ethical agent – to have "intrinsic moral status" – the entity must be *organic*. According to Torrance, only organic beings are "genuinely sentient," and only sentient beings can be "subjects of either moral concern or moral appraisal."

Some would also argue that there may not be as sharp a distinction between "explicit moral agent" and "implicit moral agent" as Moor believes, citing a neural-network approach to learning how to be ethical as falling in a gray area between the two; thus it may not be necessary that a machine be an explicit moral agent in order to be classified as an ethical machine. Others (e.g., S. L. Anderson) would say that Moor has made the correct distinction, but he has missed what is significant about the distinction from the perspective of someone who is concerned about whether machines will consistently interact with humans in an ethical fashion.

Susan Leigh Anderson makes a number of points about the field of machine ethics in "Machine Metaethics." She distinguishes between (1) building in limitations to machine behavior or requiring particular behavior of the machine according to an ideal ethical principle (or principles) that is (are) *followed by a human designer* and (2) giving *the machine* an ideal ethical principle or principles, or a learning procedure from which it can abstract the ideal principle(s), which *it* uses to guide its own behavior. In the second case – which corresponds to Moor's "explicit ethical agent" – the machine itself is reasoning on ethical matters. Creating such a machine is, in her view, the ultimate goal of machine ethics. She argues that to be accepted by the human beings with whom it interacts as being ethical, it must be able to justify its behavior by giving (an) intuitively acceptable ethical principle(s) that it has used to calculate its behavior, expressed in understandable language.

Central to the machine ethics project, Anderson maintains, is the belief (or hope) that ethics can be made computable. Anderson admits that there are still a number of ethical dilemmas in which even experts disagree about what is the right action; but she rejects Ethical Relativism, maintaining that there is agreement on many issues. She recommends that one not expect that the ethical theory, or approach to ethical theory, that one adopts be complete at this time. Because machines are created to "function in specific, limited domains," it is not necessary, she says, that the theory that is implemented have answers for every ethical dilemma. "Care should be taken," however, "to ensure that we do not permit machines to function autonomously in domains where there is controversy concerning what is correct behavior."

Unlike completeness, consistency in one's ethical beliefs, Anderson claims, "is crucial, as it is essential to rationality." Here is where "machine implementation of an ethical theory may be far superior to the average human being's attempt at following the theory," because human beings often act inconsistently when they get carried away by their emotions. A machine, on the other hand, can be programmed to rigorously follow a logically consistent principle or set of principles.

In developing ethics for a machine, one has to choose which particular theory, or approach to ethical theory, should be implemented. Anderson rejects the simple single absolute duty ethical theories that have been proposed (such as Act Utilitarianism) as all being deficient in favor of considering multiple *prima facie* duties, as W. D. Ross advocated. This approach needs to be supplemented with a decision principle to resolve conflicts that arise when the duties give conflicting advice. Although Ross didn't give us a decision principle, Anderson believes that one "could be learned by generalizing from intuitions about correct answers in particular cases."

Finally, Anderson gives a number of pragmatic reasons why it might be prudent to begin to make ethics computable by creating a program that acts as an ethical advisor to human beings before attempting to create machines that are autonomous moral agents. An even more important reason for beginning with an ethical advisor, in her view, is that one does not have to make a judgment about the status of the machine itself if it is just acting as an advisor to human beings in determining how they ought to treat other human beings. One does have to make such a judgment, she maintains, if the machine is given moral principles to follow in guiding its own behavior, because it needs to know whether it is to "count" (i.e., have moral standing) when calculating how it should behave. She believes that a judgment about the status of intelligent, autonomous ethical machines will be particularly difficult to make. (See her article, "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics," in Part IV of this volume.)

Some working in machine ethics (e.g., McClaren, Seville, and Field, whose work is included in this volume) reject Anderson's view of the ultimate goal of machine ethics, not being comfortable with permitting machines to make ethical decisions themselves. Furthermore, among those who are in agreement with her stated goal, some consider implementing different ethical theories, or approaches to ethical theory, than the prima facie duty approach that she recommends when adding an ethical dimension to machines. (See Part IV of this volume.)

J. Storrs Hall, in his article "Ethics for Machines," claims that as "computers increase in power . . . they will get smarter, more able to operate in unstructured environments, and ultimately be able to do anything a human can." He projects that they might even become more intelligent than we are. Simultaneously with their increasing abilities, the cost of such machines will come down and they will be more widely used. In this environment, regardless of whether they are conscious or not (and here he reminds us of the "problem of other minds," that we can't be certain that any other person is conscious either), "it will behoove us to have taught them well their responsibilities toward us."

Hall points out that the vast majority of people "learn moral rules by osmosis, internalizing them not unlike the rules of grammar of their native language, structuring every act as unconsciously as our inbuilt grammar structures our sentences." This learning, Hall claims, takes place because "there are structures in our brains that predispose us to learn moral codes," determining "within broad limits the kinds of codes we can learn." This latter fact explains why the moral codes of different cultures have many features in common (e.g., the ranking of rules and the ascendancy of moral rules over both common sense and self-interest), even though they may vary. The fact that we are capable of following moral rules that can conflict with self-interest demonstrates that we have evolved and flourished as social animals, accepting what is best for the group as a whole, even though it can be at odds with what is best for us as individuals.