Penultimate draft. Final version forthcoming in
Ethics and Information Technology

Moral Machines: Teaching Robots Right from Wrong by Wendell Wallach and Colin Allen[1]

Can a machine be a genuine cause of harm? The obvious answer is affirmative. The toaster that flames up and burns down a house is said to be the cause of the fire, and in some weak sense, we might even say that the toaster was responsible for it; but the toaster is *broken* or *defective*, not *immoral* and *irresponsible*, though possibly the engineer who designed it is. But what about machines that decide things before they act, that determine their own course of action? Somewhere between digital thermostats and the murderous HAL of *2001: A Space Odyssey*, autonomous machines are quickly gaining in complexity, and most certainly a day is coming when we will want to blame them for genuinely causing harm, even if philosophical issues concerning their moral status have not been fully settled. When will that be?

Without lapsing into futurology or science fiction, Wallach and Allen predict that within the next few years, "there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight" (p. 4). In this light, philosophers and engineers should not wait for a threat of robot domination before determining how to keep the behavior of machines within the scope of morality. The practical concerns to motivate such an inquiry—and indeed this book—are already here. *Moral Machines* is an introduction to this newly emerging area of *machine ethics*. It is written primarily to stimulate further inquiry by both ethicists and engineers, and as such, it does not get bogged down in dense philosophical prose or technical specification. It is, in other words, comprehensible by the general reader, who will walk away informed about why machine morality is already necessary, where we are with various attempts to implement it, and the authors' recommendations of where we need to be.

Chapter One notes the inevitable arrival of autonomous machines and the possible harm that can come from them. Some automated agents that are quickly integrating into modern life do things like regulate the power grid in the United States, monitor financial transactions, make medical diagnoses and fight on the battlefield. A failure of these systems to behave within moral parameters could have devastating consequences. As they become more and more autonomous, Wallach and Allen argue, it becomes more and more necessary that they employ "ethical subroutines" to evaluate their possible actions before they are executed.

Chapter Two notes that machine morality should unfold in the dynamic interplay between ethical sensitivity and increasingly complex autonomy, and several candidate models for automated moral agents, or AMAs, are presented. Borrowing from Moor, the authors indicate that machines can be implicitly ethical in that their behavior conforms to moral standards. Moor marks a three-fold division among kinds of ethical agents: such agents are either "implicit," "explicit" or "full". The first are constrained to emulate ethical behavior, whereas the second engage in ethical decisions making and the third are,

---

[1] This review is a slightly revised version of another that originally appeared in the January/February 2009 issue of *Philosophy Now*.

like human beings, conscious and have free will. Robots, Wallach and Allen argue, are capable of being the first, while setting the question of whether they can be explicit or full ethical agents to the side. After a brief digress in Chapter Three to address whether we really want machines making moral decisions, the issue of agency reappears in Chapter Four, where the ingredients of full moral agency (free will, understanding and consciousness) are addressed. Though machines do not currently have them and are not likely to soon, Wallach and Allen note that "functional equivalence of behavior is all that can possibly matter for the practical issues of designing AMAs" (p. 68) and that "human understanding and human consciousness emerged through biological evolution as solutions to specific challenges. They are not necessarily the only methods for meeting those challenges" (p. 69). There may be, in other words, more than one way to be a viable moral agent. The chapter ends with the provocative suggestion of a "Moral Turing Test" to evaluate the success of an AMA and the interesting suggestion that machines might actually exceed the moral capabilities of humans.

Chapter Five addresses the important matter of making ethical theory and engineering practices congruent. It also sets the stage for an important conversation on top-down and bottom-up approaches that occupies Chapters Six through Eight. A "top-down" approach is one that "takes a specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory" (p. 79-80). Rule-based systems fit here, including Asimov's "Three Laws of Robotics" and ethical theories that apply principles like utilitarianism and Kantian ethics. This is contrasted in Chapter Seven with "bottom-up" approaches that are dynamic and developmental. One might think here in terms of the moral development of a child, or, from the computer science perspective, of "genetic" algorithms that mimic natural selection to find a workable solution to a problem. Bottom-up approaches could take the form of assembling modules that allow a robot to learn from experience. Limitations that exceed the scope of this review make both of these approaches unlikely to succeed, and so, the authors recommend a hybrid approach that uses both to meet in the middle. A conventional paradigm for this approach from ethical theory is virtue ethics, where artificial agents might discover from experience "rules" that govern good behavior and use them to guide their future decisions. Such an approach would emulate the process of acquiring good character. Chapters Nine through Eleven survey existing attempts to put morality in machines, especially in light of the role that emotion and cognition play in ethical deliberation, the specific topic of Chapter Ten. The LIDA (Learning Intelligent Distribution Agent) model, based on the work of Bernard Baars, is singled out for discussion in Chapter Eleven, because it "implements ethical sensibility and reasoning capacities from more general perceptual, affective, and decision making components" (p. 172).

Finally, the closing chapter, speculates on what machine morality might mean for questions of rights, responsibility, liability, and so on. When a robot errs, who is at fault? The programmer or the robot? If it is the robot, how can we hold it accountable? If robots can be held accountable, should they also then be the recipients of rights? Not all that long ago, these questions and the whole substance of this book were the stuff of science fiction. But science fiction time and time again has become science fact. Without overstatement or alarm, Wallach and Allen make it patently clear that now is the time to consider seriously the need and the methods for teaching robots right from wrong. Of

course, attempting to understand the level of detail necessary to make robots moral and to determine what precisely we should want from them sheds considerable light on our understanding of morality in the case of human beings. In a single, thought-provoking volume, the authors not only introduce machine ethics, but also an inquiry that penetrates to the deepest foundations of ethics. The conscientious reader will, no doubt, find many challenging ideas here that will require a reassessment of her own beliefs, making this text a “must read” among recent books in philosophy and, more specifically, applied ethics.

Anthony F. Beavers, Ph.D.
Professor of Philosophy / Director of Cognitive Science
The University of Evansville
Evansville, Indiana USA