Wendell Wallach and Colin Allen: moral machines: teaching robots right from wrong

Oxford University Press, 2009, 273 pp, ISBN: 978-0-19-537404-9

Vincent Wiegel

Published online: 6 July 2010

© The Author(s) 2010. This article is published with open access at Springerlink.com

Robots and smart software have an increasing impact on our lives, and they make decisions that might have a profound effect on our welfare. Some of these decisions have a moral dimension. Hence, we need to consider whether (a) we want them making such decisions, and (b) if affirmative, how we proceed in equipping machines with “moral sensitivity” or even with “moral decision-making abilities.”

Wallach and Allen make an eloquent and forceful case that we should seriously consider granting machines such decision-making power in their book, *Moral machine, teaching robots right from wrong*. Their argument (in Chaps. 1 and 2) is that machines are deployed in situations in which they make decisions that have a moral impact. Hence we should extend them with moral sensitivity to the moral dimensions of the situations in which the increasingly autonomous machines will inevitably find themselves. This may lead to machines making moral decisions. The machines they refer to may be anything from software, softbots to robots, and in particular combinations of these. Through interconnected and open systems situations might arise that are neither desirable nor were they foreseeable when the systems were designed. Whether we can actually build such systems Chap. 3) is still an open question. If we were to engineer artificially moral systems, would they count as truly moral systems? Wallach and Allen conclude (Chap. 4) by noting that human and artificial morality will be different, but that there is no reason a priori to rule out the notion of artificial

morality. Moreover, they argue that the very attempt to construct artificial morality will prove worthwhile for all involved.

Raising these points is the first, and possibly the greatest, strength of their book. It puts the theme squarely on the agenda. Yet, theirs is also a book of open and unanswered questions. On virtually all topics, the verdict is still out, as no common opinions have been established, no approaches proven, and no answers found. Their book also serves to illustrate how young this field of research still is. Though at times it is a little disconcerting to find yet again that the answer (one of these open questions) might be A, but then again, it might not.

Writing a book that touches on several research domains—in this case moral philosophy, robotics, software developing, and neuroscience—is always a hazardous enterprise. The risk of not providing enough depth and thus losing the attention of specialists in either domain is present. The specialist will be lost unless there is enough to be learned from other domains to provide a fresh perspective on the research in his own domain.

Providing an overview of the research on artificial morality—moral philosophy and machine decision-making—is a tall order. Though the field is relatively new, there is already much and widely varied research being conducted, which ranges from moral learning algorithms and various logics to model moral decision-making, and from neural nets to nano-technology.

Overall, the book provides a good overview of most of the current research in the field, nicely setting the stage in Chap. 5 for a discussion of the relationship between engineer and philosopher; and the cooperation between the two raises various issues that occupy the remainder of the book, including questions such as: Who or what is leading? How can philosophers formulate their theories such that engineers

V. Wiegel (✉)
Section of Philosophy, Delft University of Technology,
PO Box 5015, 2600 GA Delft, The Netherlands
e-mail: v.wiegel@tudelft.nl

can actually implement them? Which moral philosophies should we use in constructing artificial moral agents?

Chapter Six discusses various top-down, rule-based approaches, whereas Chap. 7 discusses organic and emergent bottom-up approaches. Possible mixes of the two are discussed in Chap. 8. Chapter Nine contains an overview of individual research programmes currently underway. The inclusion or use of affective and emotional approaches is discussed in Chap. 10. Key issues in this chapter include concerns about the extent that emotions and social skills are necessary for moral behaviour, and about whether robots can be said to be moral if they are lacking in this respect. Chapter Eleven touches on the wider picture of artificial intelligent beings: AGIs (artificial general intelligence), also addressing some of the (dis)similarities between human and artificial morality. Turning AMAs (artificial moral agents) into beings that resemble human moral behaviour requires a much broader architectural framework, and this chapter discusses several of these frameworks and the issues associated with them.

In closing, Chap. 12 discusses futuristic scenarios and asks what our approach towards artificial moral systems might be. For example, will we (need to) assign them rights and duties? Will fear, punishment or shame have meaning in relation to robots? And what stance should politicians and legislators take towards these questions? This set of questions could be construed as providing the second main benefit or strength of the book. At times, however, the discussion remains very much at the surface. The specialist will not find anything new, while the uninitiated have too little with which to be satisfied. At best, the authors have provided a reference for where to look for further material. Yet, this is, in itself, already a benefit to book's readers.

One concern that the reader may have with this book is that the discussion, at times, gives the impression that the authors are far from fluent on the topic under discussion. This, in itself, is not surprising, and it is certainly no reproach given the range of topics. However, it can become problematic in cases where certain statements either go unreferenced or are controversial (and thus need further clarification and support). This particular flaw might easily put off some readers who happen to be better versed on the particular topic. An example is the authors' claim that "...smell and touch [...] supply information that is germane to making moral decisions." (p. 150). Also, their decision to use 'ethics' and 'morality' interchangeably without further clarification, and not to discuss the distinction between ethics and meta-ethics leaves them open to dismissal, or at least significant criticism, from moral philosophers. In a book on *Moral machines*, a few more pages could have been dedicated to making these critical distinctions more clear. Something similar also happens when the large number of modal logics is explained by the

large number of moral theories (p. 126), instead of explaining the number of modal logic models by the number of moral theories, and the number of modal logics by meta-logical considerations. These points do not affect the overall project of the book, and they might not be relevant to all of the intended readership. Nonetheless, they might distract some readers and raise some potential concerns on the representation regarding the various positions.

That the main driver for their book is not univocal or undisputed becomes clear in the first chapters in their book where they provide a fairly even-handed overview of arguments in favour and arguments against granting machines increasing decision-making powers. However, it is clear that, in their view, not considering and researching this option is dangerous because the development of technology will progress and the machine will be making (moral) decisions that affect us. One of the omissions of the book is not discussing the position of two of the most vocal opponents to their position: Johnson and Grodzinsky. One of the counter claims that Wendell and Allen's critics make is that the authors' position attributes to technological development a misleading autonomy. Critics argue that it is, in fact, we humans who drive these technological developments. And, some critics further argue that *we* (humans) can chose to stop them, or to continue, which is itself a moral decision. We can also decide not to pursue the development of technologies that have great risks attached to them.

Following a discussion of the arguments both in favour and against machines making decisions with moral aspects, the authors discuss the various ways in which engineers and moral philosophers might cooperate and approach the issue. These range from the top-down approaches, where basically rules are derived from moral theories, to leads taken from how humans 'learn' to behave morally.

The chapter on the various approaches towards actually constructing machines that make moral decisions show how broad the range of techniques and approaches is. Though the point is not explicitly made, it shows in particular how all the research projects address only a limited sub-set of aspects relevant to artificial moral decision-making. This is an important observation to be kept in mind by the researchers involved in this field.

Particularly interesting is the chapter where Wallach and Allen raise the question of how moral decision-making relates to "embodiment" and whether artificial moral decision-making should be extended with emotional and sensory components. This chapter nicely illustrates how the project of machine morality can help frame and investigate moral philosophical questions.

The subtitle of the book, *Teaching robots right from wrong*, is a catchy and beautiful sub-title. It is also somewhat misleading, however, because no-one is being taught

right from wrong. If anything emerges from the book's discussions on the topic of learning it is the impression that we still have a very long way to go. At best, we now have only a hunch about how we might go about teaching robots right from wrong. We certainly do not yet have a clear-set view on how to achieve this, and certainly there are no current "artificial pupils" capable of learning.

An important topic that is missing from the book is moral epistemology. Although it is a topic that might prove to be the hardest nut to crack, it is also one that is essential to any machine-based moral decision-making and to 'learning morality'. How is a robot to attribute moral meaning to a physical act that it perceives? Giving someone a thumb on the shoulder might be a friendly act of camaraderie by a pal from college, but also be viewed as an act of aggression when it concerns a supporter of a competing basketball team. Knowing how to interpret such actions, and next how to classify them, from a moral perspective is key to creating any kind of moral machine and thus is a theme that should be discussed more thoroughly in this book. To Wallach and Allen's credit, it is touched upon in Chap. 10, for example. But moral epistemology is a broad topic and clearly deserves explicit discussion. It is also a theme that would lead the authors back to some

meta-ethical discussions that the book avoids, such as 'Are moral properties supervenient on physical properties or are they social conventions?' These kinds of questions can have a big impact when deciding how to engineer moral decision-making. That such questions are missing from the book might be a reflection of the state of the research on machine morality.

As an introduction into the field of artificial morality and machines-based moral decision-making, *Moral machines* is a clear success. It outlines most issues pertinent to the new field, and paints a clear picture of how daunting the task of creating moral machines is. Although it might lack the depth to become the field's standard reference book, it certainly is a welcome introduction. I recommend it both as an overview of what is happening in numerous fields of research and as a quick pointer to that research. One thing that the book does well is teaching the supposed teachers of machines just how difficult their subject matter is.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.