Avant. The Journal of the Philosophical-Interdisciplinary Vanguard
Volume II, Number 1/2011 www.avant.edu.pl

Book Review: *Moral Machines. Teaching Robots Right from Wrong*

Authors: Wendell Wallach, Colin Allen
Publisher: Oxford University Press
Release date: 2008
Number of pages: 288

Dawid Lubiszewski

Translation: Ewa Bodal

Ethics and *robots* are two words that for most people remain associated primarily with science fiction movies and short stories. However, due to the rapid development of robotics over the last twenty five years, the ethical issues previously touched upon by novelists and movie directors have become subject of a scientific debate, resulting in numerous conferences, articles and studies. *Moral Machines. Teaching Robots Right from Wrong* is one of the first book publications dealing with ethical problems connected to the development of robotics. Written by two American philosophers, Professor Colin Allen from the Department of History and Philosophy of Science at Indiana University, and Wendell Wallach from Yale University's Interdisciplinary Center for Bioethics, the book points to the growing responsibilities that artificial agents, both virtual and corporeal, are charged with. As follows from this, it is necessary to develop increasingly sophisticated solutions that would allow artificial entities to make moral decisions. In their book, the authors are not afraid to pose many vital questions and to propose possible answers. Their inquiries range from whether the ethics of machines, also called roboethics, is really necessary, to the question if robots can be moral and what conditions have to be met to consider them as such. Similarly to other roboethics scholars, Allen and Wallach also emphasize that, despite the failure of the Artificial Intelligence program (namely, even though artificial entities possessing intelligence akin or identical to human have not been constructed yet), the ethical problems in contemporary robotics require a new perspective. The heretofore existing ethical systems oror the solutions thus far developed by ethicists did not involve human interactions with robots, contacts between robots and other living organisms, or interactions among robots. Allen and Wallach's book is one of the first works that consider the morality of the artificial entities as a serious matter. Consisting of 288 pages and twelve chapters in which the authors successively analyze a variety of issues, the book has been written in a language accessible not only for ethicists and engineers, but also for the people unfamiliar with the subject matter. Thus, I certainly recommend this work to anyone who wants to "be in the know" or to learn about the problems faced by roboethics, as well as the possible future of this area of studies. However, as has already been noted, the people involved with the field of study should not be disappointed by the contents of this volume either. At the very least, this claim can be supported on the basis of the following three factors: (1) the knowledge and experience of the authors regarding interdisciplinary studies; (2) the positive reviews received by the book; (3) most importantly, the very contents of the work. What is it that can be found on these 288 pages?

Most importantly, the authors present the notion of autonomous moral agents, that is, individuals aware of the moral consequences of their actions. Of course, the subject of their research is the robot rather than human, who has always been in the centre of ethical debates. Extending the definition of a moral entity to the artificial ones is not purely theoretical in application; on the contrary, it is everyday practice that has forced the creation of a new discipline of ethics. The growing presence and autonomy of artificial entities increasingly often puts them in situations requiring moral evaluation. Inability to detect such a situation and to make an appropriate response to it did, and often does cause situations dangerous for humans. According to Allen and Wallach, it is not enough to reduce the robots' capacity for wrongdoing on the level of their design. In other words, the problem does not lie in the fact that a robot's sharp parts should be covered with protective material so that it would not be able to accidentally cause anybody physical injuries. Although similar efforts might be the subject of research in engineering, their main goal should be the implementation of a system that would detect situations threatening human life that result from the robot's actions. However, the ability to undertake or desist actions, as resulting from moral evaluation, is in robotics much more problematic than it might seem, as the authors emphasize using the example of military robots, whose dynamic development and participation in military action might have been observed over the last years. The main priority of military robots is a proper execution of the ordered actions rather than moral evaluation thereof. Creating a *good* military robot and a robot that can *do* good are conflicting interests. Furthermore, the authors refer to the contemporary research, showing the extreme complexity of the human process of making moral decisions, and the challenges that robotics and artificial intelligence studies are facing today. It is not enough to implement an algorithm containing specific ethical rules; an artificial entity should also be able to put oneself in the other person's situation, a requirement that makes it much more difficult to create beings possessing abilities of moral subjects.

Conversely, one of the challenges that philosophers will have to face is abandoning general norms of behavior for guidelines appropriate in modern situations, since modern technology does not allow for creating entities that would understand such general principles as "do good" or "respect thy neighbor". Therefore, it is necessary to create very specific rules of conduct, applicable only to robots operating in a given culture and place. Another important issue raised by the authors is the subject of differences between robots and humans that may lead to significant discrepancies between the ethics developed for robots and the system developed by philosophers over more than two thousand years, as the artificial entities of today are not able to analyze the information they receive as fast as humans. As has been stated in the beginning, this book is not another work of science fiction, and thus the examples discussed therein come from contemporary technology, which may be seen as another advantage of this volume. For instance, the authors describe a case study of a moving tram, which appears in many discussions of practical ethics. Ethicists might be familiar with different variations of this problem, but the dilemma essentially consists in a choice between two possibilities. The tram may be allowed to follow one track and kill one person, or follow another track, and kill five people. Nowadays, this very problem may appear before a computer program in control of a tram. Therefore, the change may be seen as tremendous; beforehand, any occurrences of such situations could be considered individually *after* the fact by asking the human drivers why they chose the particular track. Now, however, in order to take all possible safety precautions, a computer ought to be prepared for a possibility of such an event *before* it takes place. Moreover, Allen and Wallach take into consideration the issue of the role of consciousness in morality. The authors argue that it is an excessively rigorous condition for artificial entities, since robots already exist and operate in our moral sphere, in spite of their lack of consciousness.

The authors also discuss issues connected to engineering, namely the methods of implementing an ethical system in an artificial entity. The first of the three presented methods is bottom-up, which means programming certain ready-made ethical standards. The second one is top-down, referring to the entity learning ethical standards on its own. The most promising and described in the most detail is the third method, a hybrid one which combines the two earlier propositions. To conclude, this book should be an interesting read both for a scientist and for a philosopher.