Lubiszewski, Dawid

"Moral Machines. Teaching Robots Right from Wrong", Wendell Wallach, Colin Allen, Oxford 2008 : [recenzja]

Avant 2/1, 183-189

2011

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

Recenzja książki *Moral Machines. Teaching Robots Right from Wrong*

Autorzy: Wendell Wallach, Colin Allen
Wydawnictwo: Oxford University Press
Data wydania: 2008
Ilość stron: 288

Dawid Lubiszewski

Etyka i *roboty* stanowią parę słów, która wielu osobom wciąż kojarzy się z filmami czy opowiadaniami z gatunku fantastyki naukowej. Jednakże gwałtowny rozwój robotyki w ciągu ostatniego ćwierćwiecza spowodował, że zagadnienia wcześniej poruszane przez powieściopisarzy czy reżyserów stały się przedmiotem naukowej debaty. Wynikami toczonych dyskusji są liczne konferencje i artykuły oraz badania. „Moral Machines. Teaching Robots Right from Wrong" jest jedną z pierwszych książkowych publikacji traktujących o etycznych problemach związanych z rozwojem robotyki. Napisana została ona przez dwóch amerykańskich filozofów: profesora Colina Allena z Departamentu Historii i Filozofii Nauki na Uniwersytecie Indiana oraz Wendella Wallacha z Interdyscyplinarnego Centrum Bioetycznego na Uniwersytecie Yale. Książka ta zwraca uwagę na rosnącą odpowiedzialność, jaką powierza się sztucznym agentom, zarówno wirtualnym, jak i rzeczywistym. W związku z tym konieczne jest opracowywanie coraz bardziej wyszukanych rozwiązań umożliwiających podejmowanie przez sztuczne jednostki moralnych decyzji. Autorzy nie boją się postawić w książce wielu ważnych pytań i odpowiedzi. Pytają, czy etyka maszyn, nazywana też roboetyką, jest rzeczywiście potrzebna. Czy roboty mogą być moralne i co należałoby uczynić, by można je było za takie uznać? Podobnie jak inni badacze, zajmujący się roboetyką, zwracają oni uwagę, iż mimo niepowodzenia projektu Sztucznej Inteligencji, a więc nieuzyskania sztucznych jednostek o inteligencji zbliżonej do człowieka bądź takiej samej, pojawiające się problemy z zakresu etyki we współczesnej robotyce wymaga-

ją nowego spojrzenia. Dotychczasowe systemy etyczne czy rozwiązania, jakie wypracowane zostały przez etyków, nie dotyczyły bowiem interakcji pomiędzy ludźmi a robotami, pomiędzy robotami a innymi żywymi organizmami, jak i robotami pomiędzy sobą. Praca Allena i Wallacha jest jedną z pierwszych, w której na poważnie podejmowany jest problem moralności sztucznych istot. Składa się ona z 12 rozdziałów, w których po kolei analizowane są różne zagadnienia. Całość liczy 288 stron i napisana została przystępnym językiem, zarówno dla etyków, inżynierów, jak i osób niezwiązanych na co dzień z omawianą tematyką. Jeśli więc ktoś chciałby dowiedzieć się, z jakimi problemami spotykają się roboetycy i w jaką stronę powinny iść przyszłe prace, to z pewnością powinien zajrzeć do tej książki. Jednakże, jak wcześniej zaznaczyłem, osoby związane z tematyką nie będą rozczarowane zawartością tej pozycji. Umożliwiają to przynajmniej trzy fakty. Pierwszym z nich jest wiedza i doświadczenie, jakie posiadają autorzy w badaniach interdyscyplinarnych. Drugim mogą być pozytywne recenzje, jakie otrzymała ich książka. Trzecim – najważniejszym – jest jej zawartość. Co więc znajduje się na tych 288 stronach?

Przede wszystkim autorzy przedstawiają koncepcję autonomicznie moralnych agentów, czyli jednostek, które zdają sobie sprawę z moralnych konsekwencji swoich działań. Oczywiście przedmiotem ich badań nie jest człowiek, który był zawsze w centrum dyskusji etycznych, a robot. Rozszerzenie definicji podmiotu moralnego na sztuczne jednostki nie jest jednak zabiegiem czysto teoretycznym. Jest wręcz przeciwnie, to praktyka życia codziennego wymusiła powstanie nowej etycznej dyscypliny. Rosnąca obecność i autonomia działań sztucznych jednostek powoduje, że znajdują się one coraz częściej w sytuacji wymagającej oceny moralnej. Brak możliwości wykrycia takiej sytuacji i odpowiedniej wobec niej odpowiedzi był – i niejednokrotnie jest – przyczyną stwarzania zagrożenia życia człowieka. Zdaniem autorów nie wystarczy bowiem konstrukcyjnie ograniczyć zdolności robota do czynienia zła. Innymi słowy nie chodzi o to, by robot nie zadał człowiekowi fizycznych obrażeń przy przypadkowym kontakcie z nim, ponieważ jego wystające czy ostre części pokryte zostały jakimś materiałem. Choć podobnego typu zabiegi mogą być przedmiotem prac inżynieryjnych, to ich głównym celem powinno być zaimplementowanie systemu, wykrywającego na przykład sytuacje zagrożenia życia dla człowieka, będące wynikiem działań robota. Jednak zdolność do podejmowania działań bądź ich zaniechania na skutek oceny moralnej jest o wiele bardziej problematyczna w robotyce niż mogłoby się wydawać. Pokazują to autorzy na przykładzie robotów wojskowych, których dynamiczny rozwój i udział w działaniach wojennych obserwujemy w ciągu ostatnich lat. Dla robotów wojskowych celem priorytetowym jest prawidłowe wykonanie działania a nie ocena moralna swoich czynów. Ponadto autorzy odwołując się do współcześnie prowadzonych badań pokazują jak niezmiernie złożonym jest proces podejmowania moralnych decyzji u człowieka i przed jakimi wyzwaniami staje współczesna robotyka i badania nad sztuczną inteligencją. Nie wystarczy zaimplementować w postaci algorytmu jakiś konkretnych etycznych norm, sztuczna jednostka po-

winna jeszcze umieć na przykład postawić się w sytuacji drugiej osoby. Utrudnia to znacznie stworzenie sztucznych istot posiadających cechy podmiotów moralnych.

Natomiast jednym z wyzwań, przed jakimi staną filozofowie, jest odejście od ogólnych norm postępowania na rzecz konkretnych wskazówek w pewnej sytuacji. Bowiem współczesna technologia nie pozwala na tworzenie jednostek, które zrozumiałyby ogólną zasady typu „czyń dobro", „szanuj bliźniego". Dlatego trzeba tworzyć bardzo konkretne normy postępowania, których zastosowanie ograniczać się będzie tylko do robotów działających w określonej kulturze i w określonym miejscu. Kolejnym poruszanym przez autorów ważnym zagadnieniem są różnice pomiędzy robotami a ludźmi, które powodują, że opracowywana dla robotów etyka może znacznie różnić się od tej, która opracowywana była przez filozofów przez ponad dwa tysiące lat. Sztuczne jednostki obecnie nie są w stanie tak szybko analizować docierających do nich informacji z otoczenia, jak czynią to ludzie. Tak jak napisałem na wstępie: książka ta nie jest kolejną pozycją z fantastyki naukowej. Tym samym znajdziemy w niej przykłady odwołujące się do współcześnie wykorzystywanej technologii. Jest to kolejną zaletą tej książki. Opisany zostaje między innymi pojawiający się w wielu dyskusjach z etyki praktycznej przykład z jadącym tramwajem. Etykom zapewne są znane różne wariacje tego przykładu, jednak ostatecznie dotyczą one tego samego. Mianowicie, którą z dwóch możliwości wybierzemy: czy pozwolimy tramwajowi przejechać torem pierwszym i przejechać jedną osobę, czy torem drugim i przejechać osób pięć. Współcześnie jednak przed takim problemem stanąć może program komputerowy zarządzający kolejką. Następuje więc drastyczna zmiana, przedtem bowiem tę sytuację mogliśmy rozpatrywać indywidualnie, jeśli rzeczywiście miała miejsce, i wysłuchać racji motorniczego, dlaczego zdecydował się na taki a nie inny ruch. W obecnej sytuacji powinniśmy jednak zabezpieczyć się na tyle, by program komputerowy wcześniej był przygotowany na taką ewentualność. Inna z podjętych kwestii to rola świadomości w moralności. Autorzy stoją na stanowisku, że warunek posiadania świadomości przez sztuczne jednostki jest zbyt rygorystyczny. Roboty już istnieją i działają w naszej sferze moralnej pomimo tego, iż świadomości nie posiadają.

Oprócz zagadnień filozoficznych, w książce podejmowane są też zagadnienia inżynieryjne, dotyczące metody implementacji etycznego systemu w sztuczną jednostkę. Autorzy opisują trzy metody: (a) oddolną, gdzie programuje się gotowe normy, (b) odgórną, gdzie jednostka uczy się norm, i (c) hybrydową, która jest połączeniem dwóch wcześniejszych. Ostatniej z metod poświęcają oni najwięcej miejsca i z nią wiążą największe nadzieje. Tym samym zarówno naukowiec, jak i filozof znajdą w tej książce dla siebie coś interesującego.

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 de-

cisions, 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.