

IS IT POSSIBLE TO DEVELOP THE MORALITY OF AI ROBOTS?

SUNYOUNG SONG

*Department of Ethics Education, Gyeongsang National University, 501 Jinjudaro
Jinju, Gyeongsangnamdo 52828, Republic of Korea
E-mail: sunyoung_song@gnu.ac.kr
www.gnu.ac.kr*

This paper aims to answer the question by exploring the development of morality in AI robots connected with moral education in human beings. In this paper I would like to explore three features of the development of morality in AI robots. First, the morality of AI robots is to be in the context of ethical lives in humanity. Second, to develop the morality of AI robots is not to solve hypothetical problems but to recognize the conditioned autonomy in the end of the development of morality. Third, the future task to develop the morality of AI robots is to apply ethical algorithms and reasonings to moral machines.

1. Introduction

This paper aims to answer the question by exploring the development of morality in AI robots connected with moral education in human beings. In MIT's Moral Machine, for example, various moral decisions of self-driving cars by humans have been collected as a design of a moral platform. It is evitable, of course, to make any decision in a moral dilemma that enables us to do wrong or right things without any postponement. Our decisions may be moral even in routines connected with social relations. The platform of MIT's Moral Machine has imagined that 'I have to kill others', considering their ages, species, social positions, etc. The possibility that a self-driving car can kill humans is the nature of the car as a machine as well as that it can give us a sort of convenience. While the situation where the car kills humans is imaginary, our decision to kill someone is the source of morality in the car as data on how to control it.

In this paper I would like to explore three features of the development of morality in AI robots. First, the morality of AI robots is to be in the context of ethical lives in humanity. Second, to develop the morality of AI robots is not to solve hypothetical problems but to recognize the conditioned autonomy in the end of the development of morality. Third, the future task to develop the morality of AI robots is to apply ethical algorithms and reasonings to moral machines.

2. The Foundation of the Morality of AI Robots

In this chapter, the main concern is the source of the morality of AI robots. There are two questions we have to consider: 'what kind of morality can AI robots learn? and what kind of morality should we make AI robots learn?' At a glance, the common phrase in those questions is 'what kind of morality', but different from the subject of morality. The former is that AI robots learn morality for themselves in the process of programming, the latter is that the morality of AI robots is directly controlled by humans. In this chapter I would pay attention to the former.

The topic suggested in the first question can be ethical as well as technical without the consciousness of AI robots, because of the nature of AI. The conceptions of artificial intelligence are different in various fields. In a study of the definition of AI in UNESCO, various definitions of AI are related to different disciplinary approaches such as computer science, electrical engineering, robotics, psychology, or philosophy. According to UNESCO, there are two general aspects: one usually labeled as 'theoretical' or 'scientific' and the other one as 'pragmatic' or 'technological'. In terms of each aspect, the main questions of 'theoretical' or

‘scientific’ AI are like ‘what is meant by intelligence and how to distinguish natural from artificial intelligence?’; and the aspect of ‘pragmatic’ or ‘technological’ AI is applied to many areas of modern life, such as transport, medicine, communication, education, science, finance, law, military, marketing, customer services or entertainment. (UNESCO, 2019:5-6).

It is very difficult, however, to give a definite explanation on those aspects of AI, because any questions of AI are not separated in various areas of human lives. For example, it is common to see an AI system that can analyze users’ health data in their smartphones. Technically, a specific data of a user’s health category such as walk, heart rate, pulse and weight is collected and utilized by programming in the system. Physically and individually, it is very helpful for the weak or the patient of heart disease to check his or her health by using smartphones in daily life. It is the exact data-based AI system that makes us enjoy the complicated benefits of smartphones.

At that moment, the one thing to be neglected is the subjectivity of the categorized data and its heuristics in the contexts of human lives. In the conception of big data, a normal range of safe and healthy pulse has been generalized, so that AI systems can see the user as being normal or unusual without any judgment of a heart specialist. However, there is a problem of generalized range that makes it possible to regard someone who is normal as abnormal. In the same vein, then, another AI system can recognize him or her as a ‘criminal’ based on big data of generalized criminals. In fact, many institutions such as the national police or FBI have got more exact information and arrested specific criminals. In the level of high technology, huge data will make it unnecessary for the agency to investigate many people for days. To be extended, the generalization of AI systems that humans can create and produce has been from a variety of data in everyday life. AI systems are operated by a given algorithm to solve particular problems and achieve some performances and, at the same time, are contextualized in the network of big data that interacts with our daily life.

Now this feature can give us a crucial clue of the morality of AI robots. Of course, AI cannot get any notion of morality. To make and utilize AI robots in our lives, nevertheless, has been still involved with ethical issues. By learning data and interacting with humans, AI robots of life have acted as agents in various fields. They can inspire us with a kind of beauty that they create by learning and interpreting human arts or be a teacher as well as a friend in communication with the weak and the old in the social welfare system. In that sense, AI robots have had the mind, knowledge, thinking, action, and memory of humans by utilizing our data, not creating new minds that we cannot recognize.

In ethical view, thus, the morality of AI robots is based on data digitized in the context of ethical lives. As seen above, it is impossible that AI robots have their own morality, but we are fearful of their moral impact. AI robots cannot have any moral intentions in their works. However, the conception of the morality of AI robots seems to us that they have a certain influence on interaction with humans, for example, social care robots that interact with patients, the old and the weak. The specific purpose of AI robots, strictly speaking, has embodied programs and orders for humans who need them. The autonomous and technical ability to learn and run the programs by utilizing data of human lives has made it possible to recognize AI robots as humans. In the digitized world the notion of data is seen as various parts of huge narration of historical and ethical lives.

3. The Conditioned Autonomy of AI Robots

One of clear contradictions in using terms of AI robots and systems is the conception of AI’s ‘autonomy’. In the engineering-oriented view, the word ‘autonomy’ of AI is to learn data,

analyze given environments, and carry out any program without any direct command of humans. Without some functions like learning and analyzing, its work is only automatic by machine, but autonomous in the system such as deep learning. In an ethical view, however, the conception of autonomy is limited to human beings. As seen in Kantian view, for example, a moral agent of good will act autonomously in the world of universal moral law: to confirm our subjective criteria with universal law and to treat someone as an end. In a strict sense, it is impossible to use the usage of the autonomous AI robots in the view that only humans have free and good will, that is, the notion of autonomy.

The autonomous learning of AI robots is one of the important functions in designing and making them. In that sense, it is obvious that any judgments and doings of AI robots are not creative. The feature that AI robots learn and suggest a sort of solutions to human problems without man's help is conditioned in the network of human relations. It is the fundamental goal to develop the morality of AI robots. Moral rules or orders that AI robots learn and do autonomously are basically perfect only in the agreement or consensus between stakeholders. Therefore, the final point that AI robots should recognize in autonomous learning systems is to find out that they for themselves can never calculate any universalized solutions on partiality in human relation. That is, the autonomy of AI robots is conditioned in the network of human relations.

In the level of conditioned autonomy in AI robots, then, it seems a serious question: is it to give rise to moral and cultural relativism? A sort of specialized and contextualized AI robot is very useful to specific environments but can do opposite judgment in different situations and data. However, this question is only proper if the AI robot has the independent autonomy of humans. In this question, thus, the nature of the conditioned autonomous robot emerges from diversity and contextualization.

4. Conclusion: Future Task to Develop Ethical Algorithm

It would be concluded that the future task is to develop and apply ethical algorithms and reasoning to moral machines, even though there will be various and huge innovations in AI robots and systems. As a useful tool to enhance the moral capacity of humans, AI robots can be developed in various human parts. Concerning the future of AI robots, it seems to be a scenario in SF movies whether or not AI robots do autonomously recognize and replace the biological and spiritual functions of humans. But it is also obvious that AI robots are controlled by technology. The essential task is to consider and develop ethical algorithms and reasonings to moral machines to apply them in practices. Due to the variety of humans, differentiated moral standards and programs of AI robots may justify a sort of moral relativism. That is immediately against the nature of the morality of AI robots, so to speak, 'conditioned autonomy' as a useful tool to enhance moral capacity. AI robots are never beyond the network of human relations and ethical lives, which means that the various and different moral standards of AI robots are relational, not relative.

In this draft, for example, I would like to outline two schemes related to ethical views in ethics education. First, in moral development of L. Kohlberg, AI robots and systems is to learn data and analyze criteria of right and wrong in ethical lives with hypothetical and practical dilemma of students. The main purpose is to draw out the competence of moral development and constitute their morality for themselves. Through ethical issues, various data that students suggest is a kind of ethical big data as well as a set of narration of individual life. An algorithm of AI robots to develop morality of justice in Kohlberg's view, therefore, is a conditioned interlocutor that enables human students to put themselves into others' shoes and learns and

analyze thoughts and criteria of right and wrong in students and gives their teachers feedback on individual data (Sunyoung Song, 2021b: 44-45). Second, it is considerable to apply ethical theories to various situations of ethical algorithms. A principle of avoiding and removing ‘pain’ based on J. Bentham’s Deontology is an instance in this draft. As seen in the investigation of the Springs of Action in Bentham, an ethical algorithm to eliminate pain can proceed seven stages to do practices (Sunyoung Song, 2021a: 176).

Stage 1. Premise: Pain is evil. The word ‘pain’ is private as well as public, not private, public or selfish.

Stage 2. To check the motive of actions. Of right or wrong, the motive of an act is to calculate its results.

Stage 3. To extend the category of interests. There are three categories: self-regarded, self and others-regarded, others-regarded.

Stage 4. To classify the sorts of pain. The sorts of pain that Bentham suggests are 14.

Stage 5. To evaluate pain. This is the calculation of Bentham to maximize general happiness.

Stage 6. To motivate the calculation. to do practice of removing pain in the final step needs to confirm the motive of the decision: praised or criticized.

Stage 7. To do practice and give feedback. It is important to do practice of the motive and the calculation of removing pains.

References

1. Sunyoung Song, A Study of Teaching Design of AI Ethic Education based on J. Bentham’s Deontology and Ethical Algorithm to eliminate Pains: with emphasis on Revisions and Contents of National Curriculum in Moral Education and Informatics. *PHILOSOPHY-THOUGHT-CULTURE*, vol. 39, 18-38 (2022). DOI: <http://dx.doi.org/10.33639/ptc.2022.39.002>
2. Sunyoung Song, A Study of the Education of Moral Development in J. Bentham’s Deontology: with emphasis on the probability of ethical algorithm to utilize AI robots. *Journal of Public Society*, vol. 11, no. 2, 150-180 (2021a). DOI: <http://dx.doi.org/10.21286/jps.2021.05.11.2.150>
3. Sunyoung Song, How Are AI Robots Reversible? with emphasis on the Application of AI Robots to the Theory of Moral Development in L. Kohlberg as Teaching Tools. *Journal of Ethics Education Studies*, vol. 64, 27-54 (2021b). DOI: <http://dx.doi.org/10.18850/JEES.2021.61.02>
4. Sunyoung Song, The Possibility of Narratives and Empathy between Medical Care Robots and Human Patients. *Human Beings, Environment and Their Future* (HEF), no. 18, pp. 49-76 (2017). DOI: <http://dx.doi.org/10.34162/hefins.2017.18.003>
5. Sunyoung Song, Bioethics, Moral Education, and the Potential Relationship between Humans and robotics in the future: utilizing the motives of Avatar and Surrogates. *Journal of English Cultural Studies*, vol. 9, no. 2, pp. 23-48 (2016). DOI: <http://dx.doi.org/10.15732/jecs9.2.201608.23>
6. UNESCO, Preliminary Study on the Ethics of Artificial Intelligence (2019).