The Ethical Problem of Self-driving Car

Taiga Fukuyo

Asahigaoka High School 204 19

1.Introduction

The recent development of artificial intelligence (AI), which is called the third AI boom is tremendous. Today, AI can learn on its own by using vast amounts of information. The way is called deep learning. It is the process (feature selection) of finding appropriate combination of features to focus on from some information by using the system of neural network. Taking the best behavior of professionals by using deep learning, AI can act without being taught ethics only in a very limited area such as a board game.

However, the closer AI gets to us, the more ethics it needs to be taught, because AI is active unlike other creatures and animals that have existed up to now. One AI called TAY which was invented by Microsoft in 2016 had been making discriminatory and problem statements as a result of deep learning on social media. This example might be an extreme example of involving an unspecified majority. Even so, what AI needs is not just a simple deep learning that only accumulates information, but a complex deep learning such as learning ethics in advance or learning in combined areas.

The same is true for self-driving cars. In order to achieve fully automatic driving, which is called level 5 by SAE International, it is necessary to consider behavior even in unexpected cases. However, it is not good in terms of the frame problems and various other ethical problems not to set a frame for deep learning in order to make self-driving cars behave like humans.

The purpose of this paper is to show the perspectives that will be needed in the future to make self-driving cars practical. For that purpose, I focused on problems close to the real world, which are composed of various questions described in the ethical papers so far. Various ethicists and philosophers have discussed simplified situations until now. However, almost all presented ethical problems have not yet a unified theory. The topic of this paper is like “complex system”. In a complex system, even if we observe each element individually, we cannot see the whole picture. Perhaps looking at this problem in a bird's eye view will give us a new perspective, and this concrete consideration from a developer's perspective might guide future self-driving cars developers to better way although elemental problems many ethicists have discussed have fully not been solved.

To achieve this purpose, I developed a research question, formed a hypothesis, and received opinions from a professional. As for the verification method, I firstly elected two specific currently ethical problems and consider solutions from my hypothetical viewpoint. Then I asked a professional to determine if they worked as solutions. I also

asked him other questions.

As a result, I was able to know important perspective other than the viewpoints I set. My viewpoint was to extend existing laws and rules, to teach ethics to AI, and to use the choice of the person responsible for driving to supplement laws and rules. The new perspective was that special obligations, which are determined by someone's choice should not be introduced unless everyone agrees, and that human emotions should be used as a means of teaching ethics.

A new perspective emerges by combining these results. The perspective is to expand existing laws and rules, and those that cannot be complemented are selected from well-examined legal options by drivers. AI learn ethics not only as knowledge by documents but also emotion by expressions on our faces. In general, the perspective was obtained that one method could not be compensated for and several methods should be combined.

2.Fundamentals

1) The frame problem

The frame problem is the philosophical problem in putting AI into practical use. It happens because there are too many happenings to consider for AI to act though its processing ability is limited. (Fig.1) This problem has become a barrier when self-driving cars make decisions from outside the human program. The first people to propose this problem were John McCarthy and Patrick J. Hayes (1969). Daniel Clement Dennett (1984) conducted the following thought experiment.

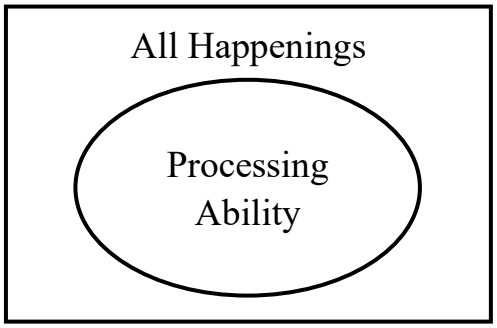

Fig.1

Inside the cave is a battery that drives the robot, on which a time bomb is mounted. The robot must get the battery out of the cave.

A) The AI robot R1 was able to take out the battery from the cave safely. However, R1 did not realize that carrying the battery would carry the bomb with it, causing the bomb to explode after exiting the cave.

B) Therefore, one have developed an AI robot R1-D1 that also takes into account incidental items. However, when the R1-D1 entered the cave and go in front of the battery, it stopped working. R1-D1 started to consider whether the ceiling will fall or not when it tries to move the bomb, whether the color of the wall will change when approaching the bomb, and whether other side effect will happen or not. Consequently, the time bomb was activated. It started thinking about everything that could occur and continued to think indefinitely.

C) Therefore, one developed an improved AI robot R2-D1 so that irrelevant matters are not considered in fulfilling the purpose. This time, however, R2-D1 stopped working before entering the cave. Before entering the cave, R2-D1 kept thinking indefinitely, trying to find out anything unrelated to the purpose. It happened because there are infinite matters unrelated to the purpose, and an infinite calculation time is required to consider all of them.

Thus, the result showed that it is difficult to make a robot R2-D2 that can judge like humans. (Daniel Clement Dennett, 1984)

This problem becomes a major obstacle in putting self-driving cars into practical use.

2)The trolley problem

The trolley problem is originally the philosophical problem of the conflict between utilitarianism and obligation theory. Philippa Ruth Foot first proposed this problem. Judith Jarvis Thomson (1986) conducted the following thought experiment.

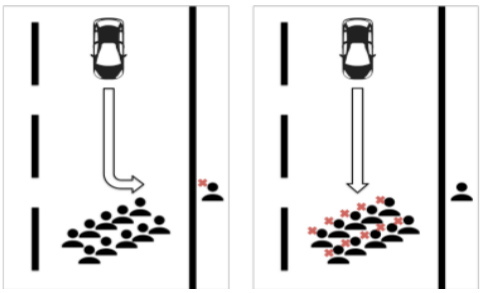

Figure 2 (ref. [6])

A) A train whose brakes have stopped working is likely to run over five workers. If a person on the junction pulls the lever, the train would change the direction and one worker is run over, and if nothing would be done, five people would be run over.

B) There is a fat person on the bridge. A train with broken brakes is likely to run over five workers. If the fat person is pushed down, the train is likely to stop.

These are often called a conflict of utilitarianism and obligation theory. From a utilitarian standpoint, they want to achieve “the greatest happiness of the greatest number” and help five workers in both situations. From an obligation perspective, a person is not allowed to kill someone for some means, and they will do nothing in both situations. However, many people tend to make different choices in these two situations. (Judith

Jarvis Thomson, 1986)

Prior research has applied this theory into self-driving cars. This is a so-called the tunnel problem. Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan (2016) took questionnaires to find out what the public wants for self-driving cars. They said "We found that participants in six Amazon Mechanical Turk studies approved of utilitarian AVs (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology." (Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, 2016: p.1573) In the paper, a dilemma happened because people wanted utilitarian cars that guarantees maximum happiness in terms of morality and might buy self-defense cars in real. This dilemma occurred in a questionnaire conducted in Japan, too. (Shigeo kawashima, et al, 2017) It may be similar to the scandal of TAY born from the gap between the real world and the real intention in terms of the gap between ideal and reality.

3.Methods

The used process

1) Developing a research question based on the purpose
2) Two specific ethical problems that are currently being discussed were accepted
3) Developing a hypothesis for the research question, and creating a procedure for solving the accepted ethical problems from the perspective of my hypothesis
4) Forming more specific problems that are likely to happen during the procedure, and considering solutions
5) A Doctor of Ethics at University of Cambridge judged whether my solutions worked as solutions and whether the concrete problems I focused on was fair as typical ethical problems. He also answered my other questions.

1) The purpose is to show the viewpoints that will be needed in the future to make self-driving cars practical. The research question to achieve that goal is “How will AI developers should deal with the ethical problems which AI face in a situation where they judge and behave themselves just like humans act under the subconscious ethical view while driving?” I developed this hypothesis because human beings, even if they do not teach ethics as knowledge, have ethics that have been formed unconsciously in the environment in which they have lived. However, this is a complex problem that we can't solve simply by letting AI behave like humans. There are countless problems to be discussed in the world. This time I considered the majority of these problems except those controlled by existing laws and rules. (Fig.3) If they are humans, each person behaves differently, so we will combine the laws later to deal with the accident. However, self-driving cars do not. The average behavior that simply extracts feature selection of human behavior by deep learning is different from individual behavior. Also, when trying to teach ethics through deep learning, there is no consistent theory on the ethical problems that are elements.

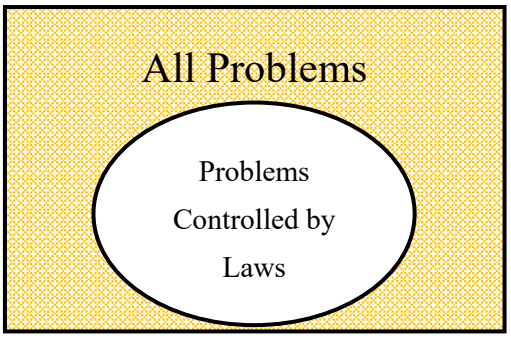

Fig.3

2) The ethical problems I accepted are the trolley problem and the frame problem. I selected the frame problem because it is a fundamental problem of self-driving cars. This problem is closely linked to other ethical problems, and if the frame problem is solved, self-driving cars that behave like humans can be created. The reason I selected the trolley problem is it is one of the most famous problem and it has largest number of previous studies among the problems related to self-driving cars. It cannot be solved even if self-driving cars that acts like humans is produced by solving the frame problem because each person has each person has their own ethics unconsciously. If I consider in the real world the frame problem and the trolley problem separately, the frame problem is too general, and the trolley problem is too specific. Therefore, in this paper, I asked a professional about the trolley problem while being aware of the frame problem. Some questions are strongly related to the frame problem.

3) My hypothesis for research questions is “Self-driving cars need to learn ethics and philosophy in advance (unlike humans) as a prerequisite for deep learning. For ethical problems that cannot be determined from a unified viewpoint such as ethics and philosophy, it is important to have a view of arranging rules and standpoints in a flexible manner by using existing laws and using the buyer’s options” Based on my hypothesis, a procedure to solve the accepted ethical problem was created. The procedure was broken down into three steps: two steps and one prerequisite step. First, in addition to existing laws, rules specific to self-driving cars that can drive more delicately than humans are added. (Fig.4) Second, in terms of morals, self-driving cars are taught ethics in advance through deep learning. (Fig.4) Finally, for ethical problems that cannot be determined from a unified viewpoint such as ethics and philosophy, it is important to have a view of arranging standpoints by using the buyer’s options. (Fig.4)

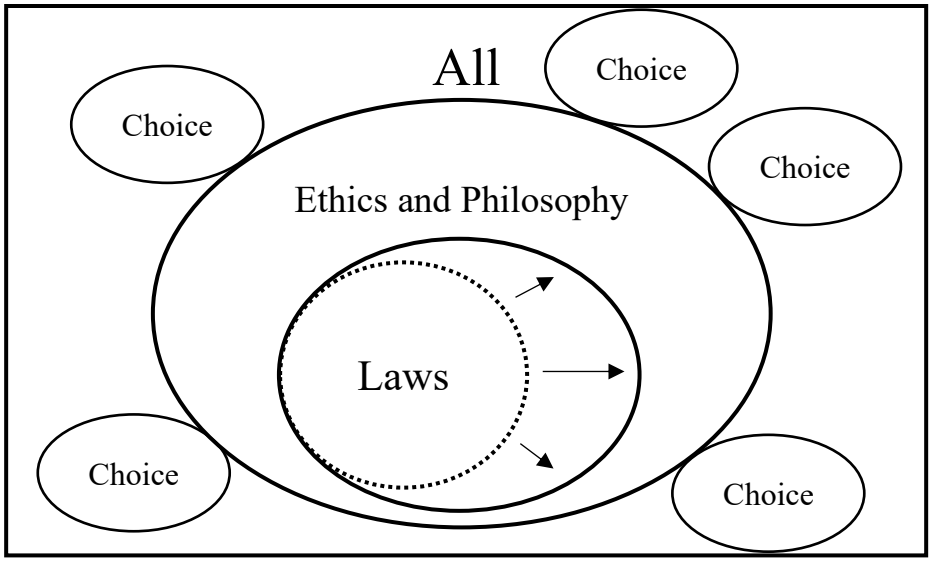

Fig.4

4) I asked questions about the trolley problem while being aware of the frame problem. I asked other questions, too.

・About extension of existing laws and rules

Q1. There are some standpoints on the trolley problem and the tunnel problem. It is difficult for each person to decide which standpoint they take, and public opinion might blame them for the choice they took when an accident happens, so should self-driving car companies put some regulations which one another?

Q2. When an accident happens, who or which should be responsible for, the drivers, car

manufacture, or the system itself?
Q3. I think if self-driving cars and cars driven by humans are on the road at the same time, something unpredictable might happen because some rules for only self-driving cars might be made. For example, the distance between two cars. Humans don't always follow the rules and AI can go out of control. In that case, who or which should be responsible?

・About the right of choice of the person responsible for self-driving

Q4. In a trolley problem, there are two different standpoints, which are utilitarianism and obligation theory. In a previous paper, when we apply this theory to self-driving cars, we might want to sell or buy utilitarian cars in ideal. However, we might buy cars in terms of obligation theory. This means we want more self-defense cars. What do you think?
Q5. I think when I develop cars, I must make it clear which standpoint I am taking. What do you think?
Q6. If a car made by a developer runs over a driver's child outside the car, it is not self-defending. Do you think that we will have such a system as we register our families when we buy self-defense cars? What do you think?

・ About teaching ethics to self-driving cars

Q7. We feel guilty when we cause an accident. However, if a self-driving car behaves according to its ethics, it does not feel guilty about their choices. Is this inevitable?
Q8. To apply ethics to self-driving cars, deep learning like facial recognition systems is effective, I think. Do you have any idea we should think carefully in terms of ethics and what and how we should regard as more important?

5) I met and talked with Dr. Simon Beard, who works at Center for the Study of Existential Risk in University of Cambridge. (1 AUG 2019) He works on the project "Managing Extreme Technological Risks and mainly studies on consequentialist moral philosophy."

4.Results

・About extension of existing laws and rules

Q1. There are some standpoints on the trolley problem and the tunnel problem. It is difficult for each person to decide which standpoint they take, and public opinion might blame them for the choice they took when an accident happens, so should self-driving car companies put some regulations which one another?

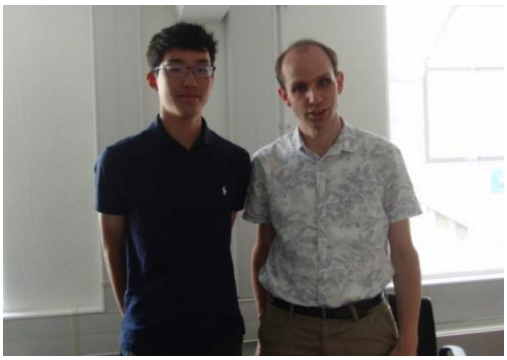

Fig.5 (Dr. Simon Beard (right) and me)

A1. This is difficult one. The company is the group of people and each person has a different role in the company. It's hard to get all these people to take the same ethical standpoint. I think it is better if everyone can agree on the ethical standpoint but it's very hard.

Q2. When an accident happens, who or which should be responsible for, the drivers, car manufacture, or the system itself?

A2. This is probably the biggest problem which is going to prevent this technology. I think this is a legal problem rather than a philosophical problem. They need to make clear rules by the courts. The best way to apportion responsibility is going to be to whoever can make it safer. If a car hits someone, it's already done something unsafe. Ideally, the cars will never hit anyone, and they never have to make ethical choices.

Q3. I think if self-driving cars and cars driven by humans are on the road at the same time, something unpredictable might happen because some rules for only self-driving cars might be made. For example, the distance between two cars. Humans don't always follow the rules and AI can go out of control. In that case, who or which should be responsible?

A3. One important thing is there are also rules drivers must follow. If you are following the rules on the highway course, you are being a good driver. The Al cars must follow lots more ethical rules and technical rules. I think if the human driver did something more dangerous but what they did was still in their rules and the AI car did less dangerous, but it broke one of their rules, we might say that the AI was responsible. (Fig.6) If both observe their own rules and the human breaks the AI car rules, no one would be responsible. (Fig.7)

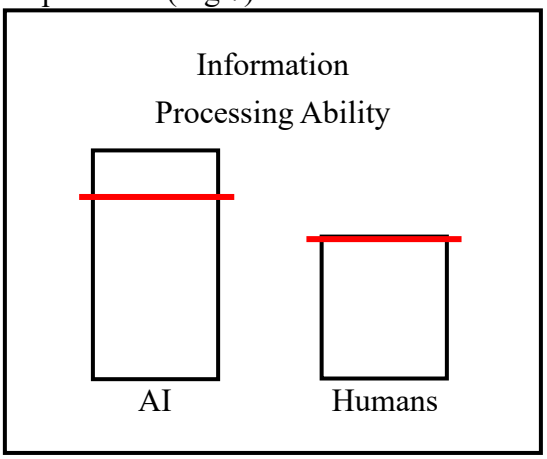

Fig.6

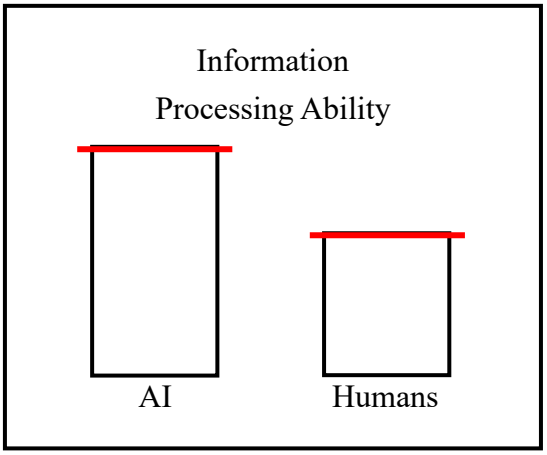

Fig.7

・About the right of choice of the person responsible for self-driving

Q4. In a trolley problem, there are two different standpoints, which are utilitarianism and obligation theory. In a previous paper, when we apply this theory to self-driving cars, we might want to sell or buy utilitarian cars in ideal. However, we might buy cars in terms of obligation theory. This means we want more self-defense cars. What do you think?

A4. I think most philosophers find it easy to say that you should turn the trolley and you should not kill the five people. If the one person is you, it seems more ethically important that you accept this version in order to save more people. People are not philosophers. The standpoint of the problem seems to be less different between the moral theories and more different between behaviors as philosophers would like to behave or ordinary people who don’t think about the role of ethical terms would like to behave.

Q5. I think when I develop cars, I must make it clear which standpoint I am taking. What do you think?

A5. Yes, I think that's really important. I think one of the mistakes that some people make when they develop cars is that they have the only standpoint they can think about as if they are the drivers, as if they are the AI, but If you saw a car crash happening and you saw what the driver did, what you would think is about the driver's actions. It's better here for them to be able to take a different standpoint that makes difference.

Q6. If a car made by a developer runs over a driver's child outside the car, it is not self-defending. Do you think that we will have such a system as we register our families when we buy self-defense cars? What do you think?

A6. I don't think that's a good idea. I think if you want to give the car special obligations to protect someone, you should give car obligations that everyone agrees to, rather than the person who buys the car. We need to have rules about who should be protected. If there is a rule that we want, every car should have it.

・ About teaching ethics to self-driving cars

Q7. We feel guilty when we cause an accident. However, if a self-driving car behaves according to its ethics, it does not feel guilty about their choices. Is this inevitable?

A7. This is something that worries me a lot in the way the people about AI. There are something missing in the philosophical ethics which doesn't tend to talk about emotions. I think need to find the way to welcome AI into our communities to sympathize with AI and understand what the AI is. One important thing about the car is "learn". We need to be able to tell the car "I feel bad about what you chose." by machine learning. I think that can be very important part of how society develops to accept the self-driving cars.

Q8. To apply ethics to self-driving cars, deep learning like facial recognition systems is effective, I think. Do you have any idea we should think carefully in terms of ethics and what and how we should regard as more important?

A8. One idea I would like to see people develop is using the machine learning of the reinforcement learning for cars to work out for themselves what people want them to do. What we could do is to tell the cars in the simulator how people watch what the cars do, and they could say it is the right thing or it is the wrong thing. The car could learn for itself what people's rules are.

5.Discussion

First, I write about the expansion of laws and rules. When we lay new special rules and duties down, all people should agree and apply the same rules and duties in all cases. This situation is very severe, because a community is a group of people and each of them plays a different role. It is difficult for everyone to take the same position naturally. Therefore, it should be set in the form of a law that is carefully examined by courts.

Next, I write about a situation that cannot be determined by the expanded law. This is a situation in which each person behaves differently and later looks back on the behave by law. It is generally a situation where an unexpected accident has occurred. The developer indicates to the public what ethics he or she developed. As a result, the purchaser can select self-driving car's behavior at the time of purchase in advance. The difference of the selection is not caused by a theoretical difference in ethical problems but caused by rather an acceptance of exceptions in applying it to the real world.

Finally, I will write about ethics and philosophy education for self-driving cars. Self-driving cars feel potentially less guilty than humans. Also, theoretical philosophy as a discipline tends not to consider emotions. One of the solutions to this situation is machine learning. This method with machine learning not only gives a huge amount of knowledge on paper to AI, but also uses a simulator to indirectly share human emotions as human expressions and impressions on the behavior of self-driving cars.

The procedures based on my hypothesis are compared to these results.
[About first step] Dr. Simon Beard agreed and gave me a concrete method. (From A.3)
[About second step] He taught me to use human emotions as well as paper knowledge as a way to teach ethics. (From A.7) One reason is that this dilemma is not a conflict of ethical theories, but a dilemma between philosophers and people. (From A.4) It is difficult for us to unify which to teach in ethics knowledge on paper to AI even within one group (From A.1), but at least clarify the standpoint is necessary. (From A.5) He also showed me concrete method using a simulator as an example. (From A.8)
[About second step] He told me it was not a very good way. He also said that if you wanted to have AI have a special obligation, you should establish a law that everyone agrees with. (From A.6)
[About the whole] He said that the main premise is that engineers make things safe.

He also said that many are legal rather than ethical arguments. He also said that much of this discussion was about the law rather than about ethics. (From A.2)

6.Conclusions

This time, I considered the trolley problem and the frame problem as typical examples of the ethical problems. The keywords in this research are "emotion" and "law." There is a simulator for teaching emotions. In January 2020, TOYATA announced a plan for “Woven City”. This system is very useful as a bridge between the simulator and real life. I’m looking forward to the future of AI.

I have loved creating things since I was little, and in the future I want to work in engineering. In particular, I want to be a developer who uses AI in real life. Therefore, I thought it was important for developers to understand the ethical problems involved in development, so I researched it. Through this paper, I am also interested in the laws related to development and think that it is important for development now, so I would like to look at the laws next. He also said that it's best for developers to make safe things, and I think so, too. In any case, if a developer make a perfect one, the problems related to it will not occur. I looked back at "development" in a different way than usual, and realized the importance of it again, too.

7.Acknowledgements

I am deeply grateful to Dr. Simon Beard for all his help. He gave me his valuable time and opinions about my questions. He was kind enough to respond politely to my poor English and to talk with me for an hour. Not all of them can be written, and some of the written content may be different from his comments by my mistakes. I also thank Ms. Kurokawa, Mr. Tanaka, Ms. Nakamura and Mr. Narita in my high school and Mr. Suzuki and Mr. Yasuhara who are graduates of my school. They supported and encouraged me at any time.

8.Reference

[1] Bonnefon, J., Shariff, A., and Rahwan, I. (2016) *The social dilemma of autonomous vehicles*. Science. vol.352, p.1573-1576.

[2] Dennett, D. C. (1984) *Cognitive Wheels: The Frame Problem of AI.* Cambridge. Cambridge University Press. p.147-170.

[3] Foot, P. (1978) *The Problem of Abortion and the Doctrine of the Double Effect* in *Virtues and Vices.* Oxford Basil Blackwell.

[4] Japanese Engadget. (2016) *Microsoft apologized for Tay's discriminatory remarks. They described it as "systematic attacks exploiting vulnerabilities".* [Online] Available at: https://japanese.engadget.com/2016/03/28/tay/,(accessed 18 July 2019).

[5] Japanese Nursing Association. (2019) *What is ethics?* [Online] Available at: https://www.nurse.or.jp/nursing/practice/rinri/text/basic/what_is/index.html, (accessed 17 July 2019).

[6] Kawashima, S., et al. (2017) *Consciousness concerning "the trolley problems" of automatic driving vehicles*: *The results of a questionnaire survey of people living in Japan* [Online] Available at: http://gmshattori.komazawa-u.ac.jp/ssi2017/wp-content/uploads/2017/07/10.pdf, (accessed 18 July 2019)

[7] Matsuo, Y. (2015) *Will artificial intelligence go beyond humans? What lies ahead of deep learning. Japan.* KADOKAWA Corp.

[8] McCarthy, J. and Hayes, P. J. (1969) *Some philosophical problems from the standpoint of artificial intelligence*. Machine Intelligence 4, p.463-502.

[9] News Pics. (2016) *Autonomous driving research by Toyota: Could they use AI to respond to unexpected situations?* [Online] Available at: https://newspicks.com/news/1358347/body/, (accessed 17 July 2019)

[10] Okamoto, Y. (2018) *If we teach AI philosophy? Japan.* SB Creative Corp.

[11] Self-driving Lab./ (2019) *Summary of basic knowledge of autonomous driving levels (Latest edition)* [Online] Available at: https://jidounten-lab.com/u_1057, (accessed 17 July 2019)

[12] Shibata, M. (2001) *Heart of robot: The Seven Philosophical Stories. Japan.* Kodansha Ltd.