Kseniia Zobenko
Odesa I. I. Mechnikov National University, Ukraine

ABOUT FAT MEN, TROLLEYS AND MORAL PHILOSOPHY OF ROBOTS

The article presents an analysis of Philippe Foot's mental experiment a 'trolley problem', which was intended to show the differences between intention and foresight in moral action; and analyzes the ethics of robots in the first season of the series "The WestWorld". The specificity of the philosophy of morality of a person is shown as a result of the analysis. The question is also posed: what problems can arise when a person undertakes to create a completely new thinking and feeling being? The "WestWorld" shows hypothetical consequences (maybe hyperbolic), if a person created a robot with a stronger emotional attachment to this world. Such science fiction as a thought experiment exists specifically to depict the dark scenarios of our future, so that we do not have to live it.
Keywords: moral philosophy, ethics, trolley problem, mental experiment, robots, artificial intelligence.

Introduction. The concept of artificial intelligence did not appear suddenly – this is the subject of deep philosophical discussions that are most discussed and criticized today: can a machine really think like a person? Can a machine be human? The French physicist, mathematician and philosopher – Rene Descartes – was one of the first to be interested in these problems, and in 1637 in his book "Discourses on the Method" he reflects on these issues. Descartes actually summarizes some of the key points and challenges that technology needs to overcome. Rene Descartes, in her reflections on the peculiarities of the differences in the thinking and structure of man and animals, suggested that the animals have a certain complex mechanism. But provided we create a machine endowed with the same bodies and appearance as, for example, a monkey or other animal, we will not have any way of knowing the same breed as this animal. In the event, however, if we create a machine that is similar to a human body and endowed with human actions, then Descartes assumes that we would have two ways to know that we are not human beings before us. First (in Descartes' representations, which were based on the knowledge of the XVII century), such a machine would not be able to express its thoughts in words, like a person, and if, nevertheless, he could not handle the logical answers in his presence. (Now we know that cognitive work is able to simulate even emotions, feelings, and cunning and cunning). Secondly, an artificially created machine similar to a person might have performed many tasks more quickly and effectively than a person, but it was not deliberate, but only because its bodies are assigned to perform those tasks and actions.

And the emergence in the late 40-ies of XX century electronic digital computers, possessing universal capabilities and high productivity – immediately gave rise to the question: can machines of this type (with their further improvement) "think" like that person? In other words, is it possible to create a machine whose intellectual capabilities would be identical to the intellectual capabilities of a person (or even surpass human capabilities)?

One of the most promising and criticized areas of research in robotics is associated with the creation of machines capable of acting as moral subjects. In other words, having the opportunity to commit acts in a situation of moral dilemmas. These developments are necessary, especially in modern conditions, when the mass distribution of unmanned vehicles, nurse care bots and other devices is near, the interaction with which directly affects human life, and which are often forced to react quickly to new circumstances.[1] In fact, robots must be taught to work with the "trolley problem", which has ceased to be only a mental experiment and turned into a challenge to neuroethics and robotics. Scientists suggest that since we are talking about the need to make rational decisions (even if they are moral), robots that can learn in the process of applying solutions to a multitude of problems (say, learn from precedents) can generally cope

[1] Sachdeva, S. (2011). Culture and the quest for Universal principles in moral reasoning. *International Journal of Psychology*, 3, 161-176.

with this task.[1] For example, you can offer a computing system to analyze a huge number of cases so that in the end it will form some averaged ethical recommendations. In this case, the machine cannot be a fully independent moral agent, but it will be impossible to accuse it of inhumanity: its actions will be dictated by the experience and habits of people relevant to a given situation. Maybe the machine will not be able to survive certain sensations, but it will be able to express them – and in fact, we traditionally consider the ability to correctly represent certain decisions and emotions as a sign of a socialized, "normal" individual.

Sometimes in science fiction there is material that can be viewed as a philosophical experiment that examines hypothetical situations that model possible probabilities: as the American philosopher Susan Schneider[2] notes, "thought experiments are imaginations of the imagination; they are windows to the fundamental nature of things. A philosophical mental experiment is a hypothetical situation in a "mind laboratory" that portrays something that often exceeds the limits of modern technology or is even incompatible with the laws of nature, but it should reveal something philosophically enlightening or fundamental to the topic in question. Experiments with thinking can demonstrate a point, entertain, illustrate a riddle, expose a contradiction in thoughts and force us to give an additional explanation".[3] So the stories of science fiction writer Isaac Asimov were prerequisites for explaining and representing the morality of robots. From the story "Runaround" (1942)[4] Three Laws of Robotics were derived. They can be described as moral and ethical:

11. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
12. A robot must obey the orders given it by human beings except where such orders would conflict with the first law.
13. A robot must protect its own existence as long as such protection does not conflict with the first or second laws.

However, many of Asimov's stories illustrated the unintended problems that could occur as a result of following these rules. The rules are of course fictional, and there is no simple way of translating them into implementable code. What should be the program to make sure that the action or inaction of the robot does not lead to what harm the person? Will the robot be able to foresee all the possible consequences of their actions and their interaction with human behavior? The rules are more focused on short-term physical security, when, obviously, there are other problems that people may suffer. Robot actions can also lead to other types of damage, such as psychological trauma or emotional distress. The rules imply robots that can understand the orders given to them by humans (and the extent to which they contradict the first law).

The difference between mental experiments in philosophy and science fiction is in solving the problem: the task of philosophical thought – the experiment as a whole; and the task of science fiction is entertainment in general. Exploring the ethical problems of abortion, Philippa Foote created a mental experiment "trolley problem", in her 1967 paper, «Abortion and the Doctrine of Double Effect»[5] which was primarily intended to clearly show the differences between intention and foresight in a moral act, due to the possible death of a person (it turned out that in any moral deed). You see a runaway trolley moving toward five tied-up (or otherwise incapacitated) people lying on the tracks. You are standing next to a lever that controls a switch. If you pull the lever, the trolley will be redirected onto a side track and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options: 1) do nothing and allow the trolley to kill the five people on the main track; 2) pull the lever, diverting the trolley onto the side track where it will kill one person. Which is a more ethical option?

In an attempt to resolve this dilemma within the framework of moral philosophy, there is a confrontation between each other the concept of utilitarianism (J. Stuart Mill, Jeremy Bentham, and

[1] Friesdorf, R., Conway P., Gawronski B. (2015). Gender differences in response to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin*, 41, 696-713.
[2] Schneider, S. (2009). Introduction: Thought Experiments: Science Fiction as a Window into Philosophical Puzzles. *Science Fiction and Philosophy: From Time Travel to Superintelligence*. Hoboken: Wiley-Blackwell, 1–14.
[3] Ibid, 1–14.
[4] Asimov, I. (1942). *Runaround. Astounding Science Fiction*. New York: Street & Smith, 78-94.
[5] Foot, P. (1978). *The problem of Abortion and the Doctrine of the Double Effect, in Virtues and Vices and Other Essays in Moral Philosophy*. Berkeley: University of California Press, 207.

others) and deontological ethics that inherit the ideas of Immanuel Kant. According to utilitarians, those actions that increase the general benefit and happiness are preferable – it means that you need to switch the arrow without any questions. But from the point of view of Kant's ethics it is impossible to do this. A person can only be seen as a goal, not a means, even if it is a means to save the lives of others.

In 1976, Judith Thomson[1] proposed an improved version of the experiment, typing a fat man into the formulation. The wording "Fat man": "As before, the trolley is carried along rails to which five people are tied. You are on a bridge that passes over the rails. You have the opportunity to stop the trolley, throwing something heavy on the way. Next to you is a fat man, and the only way to stop the cart is to push it off the bridge on the way. You can throw yourself under the train, but your weight will not be enough to stop the cart. What are your actions? "[2] Choosing the right solution here becomes much more difficult. From the point of view of the pure utilitarian, nothing has changed: you are also morally obligated to sacrifice one in order to save five, even though this is connected with a direct violent act. But most people, as shown by studies and surveys conducted in several countries, I think differently. If in the first variant of the experiment almost everyone is ready to switch the arrow, then the number of those who want to kill the fat man is much smaller. What is the matter? What exactly has changed in the second experiment compared to the first? Many ways have already been proposed to explain the difference between the case of a change in the arrow and the case of a fat man. For example, it can be noted that in the second situation a fat man can survive, while in the first one someone will surely die. If in the first case you send the arrow, then you can confidently and confidently, you will, of course, be very happy – your conscience will not be burdened with anything. But if a fat man can be saved and does not stop him with his body, there will be nothing joyful in this: it turns out that you did not save anyone and made an inadvertent attempted murder.

Psychologist Joshua Green of Harvard University and J. Cohen used the methods of functional magnetic resonance imaging to study the problem of the trolley. In their experiments, people's answers were analyzed to the questions posed in the original formulation and in the wording with the "fat man." The scientist's hypothesis was that the solution to these problems will cause both emotional and cognitive reaction, and their conflict will arise. The results of the study showed the following: in situations that cause an emotional response ("fat man"), significant activity is observed in those parts of the brain that are associated with resolving the conflict. At the same time, in more neutral situations (for example, the original trolley problem), activity is observed in the brain area responsible for higher cognitive functions. Joshua Green argues that in the first case we are dealing with an impersonal moral dilemma in which cold rational thinking is involved. And in the second case, emotions and empathy are included, which lead to completely different solutions. This explanation looks very plausible, especially when you consider that a fat man is much more likely to kill autists and people with psychopathic inclinations who have great problems with empathy. Thus, potential ethical ideas in a given situation revolve around a person's ability to make rational decisions of moral character.

Another interesting example of solving the dilemma inside the trolley problem was the assumption that introducing uncertainty in the conditions changes how people think about the dilemma of the trolley as a whole. In a 2014 study, Kathryn Kortenkamp and Colleen F. Moore gave subjects two versions of the trolley problem for the first time: the first where the results that were to occur were guaranteed; and the second, where there were no guarantees.[3] In the latter case, respondents were less inclined to think that killing one person for the sake of saving five was "acceptable" or "moral." In the face of uncertainty, "participants may have relied more on deontological," or reason-based reasoning, "than on utilitarian moral judgments," say the researchers. The latter depends on the ability to know what will be the result of your actions, but if you are not sure, you may feel that acting in accordance with moral teachings is not only ethically correct but more safely. Kortenkamp and Moore: "These data suggest that in the study of moral judgments and reasoning regarding uncertain situations, the fact of uncertainty must be taken into account."[4]

[1] Thomson, J.J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist*. Oxford: Oxford University Press, 59, 2, 204-217.
[2] Thomson, J.J. (1985). The Trolley Problem. *The Yale Law Journal*, 94, 6, 1395-1415.
[3] Kortenkamp, K.V., Moore, C.F. (2014). Ethics Under Uncertainty: The Morality and Appropriateness of Utilitarianism When Outcomes Are Uncertain. *The American Journal of Psychology*, 127, 3, 367-382.
[4] Kortenkamp, K.V., Moore, C.F. (2014). Ethics Under Uncertainty: The Morality and Appropriateness of Utilitarianism When Outcomes Are Uncertain. *The American Journal of Psychology*, 127, 3, 367-382.

Will robots change our morality?

Of course, automatic machines do not program themselves. They will be programmed by people who are aware of the consequences of various algorithms. Thus, these programs intentionally aim at a specific result. But during the operation, automatic machines make decisions. This is one of the properties of artificial intelligence. At this stage, the program is removed from the process, and the more artificial intelligence is developed, the more computers acquire the ability to learn independently, the more distant from the solutions is the program. You see, the mechanisms will make decisions that are not even predicted in their program. One of the common problems arising in modern sciences, philosophy and science fiction is the expansion of human intellectual and physical abilities and capabilities with the help of human improvement technologies and artificial intelligence. Expansion involves not only technological solutions but also ethical, philosophical, sociological and other decisions because of the fears of advanced technologies that can replace a person or simply destroy him.

Some science fiction works are devoted to the question of expanding the intellectual and physical abilities and capabilities of a person with the help of technologies of human improvement and artificial intelligence and the fears generated by this problem. One of the latest striking examples of such science fiction works is the American series "The WestWorld". At the moment, in the real world, our robots can only think mechanically and they can only depict emotions and only depict that they understand your emotions. Your digital friend may sound like it's really concerned about your question and will try its best to answer it, but we all know that there's nothing behind it. «WestWorld» asks us to think about what could happen if we could build a robot with a stronger emotional attachment to this world. He also studies all the problems that we can face if we create robots that look like humans but use them to satisfy our darkest desires. For example, one of the main characters of the "Western world" – Robert Ford, plays God, who controls the world, which he created himself, and raises nietzschean deism, in which God creates the world and its laws of existence, and then killed by his creations.

Humans should not weaken their responsibility for the consequences of the actions of the robot that perform them. It is also important, recognizing this responsibility, to attempt to predict the potential negative consequences of engaging robots in situations where moral decisions are required, and to make efforts to limit their use. Respectively designed robots can bring many benefits to human society, but a responsible approach to robotics should be designed to limit their incursions into morally sensitive situations before it is too late.

Conclusion. As the philosopher Vilém Flusser wrote, history knows several industrial revolutions, but today we are witnessing a qualitatively new round of relations between man and technology.[1] One of the most promising and criticized areas of research in robotics is associated with the creation of machines capable of acting as moral subjects. In other words, having the ability to commit acts in a situation of moral dilemmas. These developments are necessary, especially in modern conditions, when the mass distribution of unmanned vehicles, nurse care bots and other devices is near, and human life is directly linked to interaction with them. In fact, robots need to be taught how to work with the "trolley problem". The "trolley problem" ceased to be only a mental experiment and turned into a challenge to neuroethics and robotics. So, even without being able to construct an electronic duplicate of the will, a person still inevitably encounters the need to calibrate the ethics of social interactions.

The "Westworld" gives impetus to the idea of justice and morality of the existing systems of actions; to the idea of what exactly they can "teach"; and how to reproduce in one's own creations that which is intuitively or quite frankly considered ethically ambivalent.

David Edmunds believes that humans would prefer to program robots to minimize victims and maximize benefits.[2] And it would be quite reasonable: if the person's emotionality prevents him from killing the fat man, then the machines have no such excuse. Their ethics will be different from ours. But it is our arguments, experiments and reasoning that in many respects will determine exactly what it will be.

Rene Descartes, in her reflections on the peculiarities of the differences in the thinking and structure of man and animals, suggested that the animals have a certain complex mechanism. But provided we create a car endowed with the same bodies and appearance as, for example, a monkey or other animal, we will not have any way of knowing the same breed as this animal. In the event, however, if we create a machine that

[1] Flusser, V. (2018). *Vom Stand Der Dinge. Eine Kleine Philosophie Des Design.* Verlag: Steidl Gerhard, 160.

[2] Edmonds, D. (2015). *Would You Kill a Fat man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong.* Princeton: Princeton University Press, 240.

is similar to a human body and endowed with human actions, then Descartes assumes that we would have two ways to know that we are not human beings before us. First (in Descartes' representations, which were based on the knowledge of the XVII century), such a machine would not be able to express its thoughts in words, like a person, and if, nevertheless, he could not handle the logical answers in his presence. (Now we know that cognitive work is able to simulate even emotions, feelings, and cunning and cunning). Secondly, an artificially created machine similar to a person might have performed many tasks more quickly and effectively than a person, but it was not deliberate, but only because its bodies are assigned to perform those tasks and actions.

References:

1. Asimov, I. (1942). *Runaround. Astounding Science Fiction*. New York: Street & Smith. [in English].
2. Edmonds, D. (2015). *Would You Kill a Fat man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong.* Princeton: Princeton University Press. [in English].
3. Friesdorf, R., Conway P., Gawronski B. (2015). Gender differences in response to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin*, 41, 696-713. [in English].
4. Thomson, J.J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist*. Oxford: Oxford University Press, 59, 2, 204-217. [in English].
5. Thomson, J.J. (1985). The Trolley Problem. *The Yale Law Journal*, 94, 6, 1395-1415. [in English].
6. Kortenkamp, K.V., Moore, C.F. (2014). Ethics Under Uncertainty: The Morality and Appropriateness of Utilitarianism When Outcomes Are Uncertain. *The American Journal of Psychology*, 127, 3, 367-382. [in English].
7. Foot, P. (1978). *The problem of Abortion and the Doctrine of the Double Effect, in Virtues and Vices and Other Essays in Moral Philosophy*. Berkeley: University of California Press. [in English].
8. Sachdeva, S. (2011). Culture and the quest for Universal principles in moral reasoning. *International Journal of Psychology*, 3, 161-176. [in English].
9. Schneider, S. (2009). Introduction: Thought Experiments: Science Fiction as a Window into Philosophical Puzzles. *Science Fiction and Philosophy: From Time Travel to Superintelligence*. Hoboken: Wiley-Blackwell, 1–14. [in English].
10. Flusser, V. (2018). *Vom Stand Der Dinge. Eine Kleine Philosophie Des Design.* [From the state of things. A small philosophy of design]. Verlag: Steidl Gerhard. [in Deutsch].