ГБОУ ГИМНАЗИЯ № 248 САНКТ-ПЕТЕРБУРГА

Индивидуальный конкурс эссе

“Единство в различии”

Исследовательская работа (эссе)

«Робоэтика (по законам Азимова) и человеческая мораль: где грань допустимого? »

Выполнила: Андреева Виктория Юрьевна,

Ученица 10 класса

Руководители:

Камитова А.В.

Шляховская Л.Н.

2023

Saint Petersburg Gymnasium (Grammar School) № 248

Individual essay contest

“Unity in Diversity: Russia and the English-Speaking World. Time for Equal Opportunities”

Research work (essay)

Robot ethics (according to Asimov’s Laws) and moral dilemmas: where is the limit?

Written by: Andreeva Viktoria

Supervisors:

Kamitova A., Shlyakhovskaya L.

2023

Theses

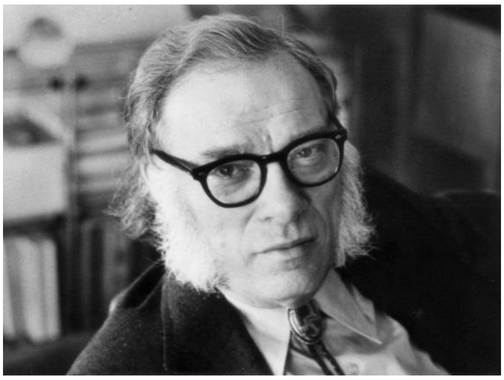
1. The scope of human robot interactions is growing.
2. Asimov’s Three Laws of Robotics.
3. The moral justifications of the Three Laws.
4. Moral dilemmas - situations of choice in which someone is faced with a moral choice between two possibilities.
5. The AI does not have algorithms for acting in case of all unforeseen situations.
6. The line separating robot ethics and morality.

Robot ethics (according to Asimov's Laws) and moral dilemmas: where is the limit?

As robotics technology becomes more ubiquitous, the scope of human robot interactions is growing. The main reason for the question about the ethical behaviour of robots is the rapid progress in the development of autonomous social robots that are specifically created to be deployed in sensitive human environments, health care settings and military contexts. Clearly, such tasks and environments are very different from traditional factory environments. These new social robots will require higher degrees of capacity for independent, self-directed and decision-making actions than any previously developed machines, given that they will face a much more complex, open world.

It is crucial to understand whether robots can act in difficult situations, whether they can make the right moral choice in the so-called "moral dilemmas", and where the line between robot ethics and morality is.

In 1942 in a short story 'Runaround', the science-fiction writer Isaac Asimov introduced the Three Laws of Robotics — engineering safeguards and built-in ethical principles that he would go on to use in dozens of stories and novels. They were:



Isaac Asimov (1920-1992)

"1) A robot may not injure a human being or, through inaction, allow a human being to come to harm;

2) A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law;

3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws."[1]

---

[1] Isaac Asimov, Runaround (1942, p.27)

Based on these laws, people have been creating artificial intelligence (AI) for many years. These laws are closely related to the basic laws of morality, which can be seen in different religions.

In the story "Evidence" (1946), Asimov described in detail the moral justifications of the Three Laws. The heroine of the story, Dr. Susan Kelvin, gives the following arguments:

“A person usually refrains from another person, except in cases of acute coercion (for example, in war) or to save more people. This is the equivalent to the First Law.

Similarly, feeling responsible to society, a person follows the instructions of authoritative people: doctors, teachers, bosses, etc., which corresponds to the Second Law.

Finally, each of us cares about our own safety — and this is the Third Law.”[2]

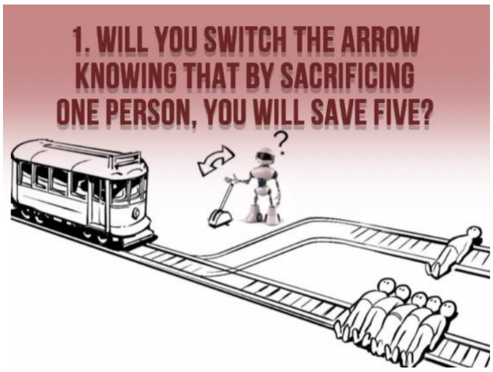
Thereby, the laws of robotics correspond to morality as much as possible. Generally accepted traditions, based on the concepts of good and evil, are aimed at the benefit of humanity.

There are moral dilemmas - situations of choice in which someone is faced with a moral choice between two possibilities, in which the choice of either of them is associated with a violation of certain moral prescriptions. Discussions about whether it is possible to fully comply with the requirements of morality and not to commit acts contrary to these norms have been going on for many centuries.

In all examples except the last one there is a humanoid robot that can do certain actions.

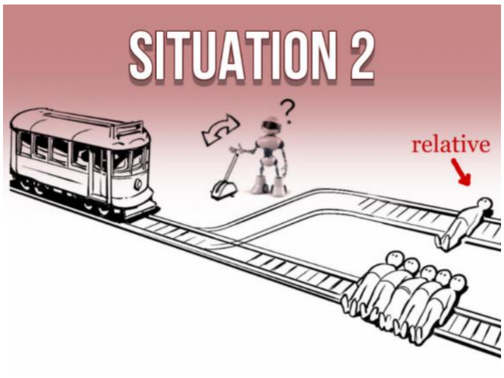
[2] Issac Asimov, Evidence (1946,p.121)

Let us consider the most famous example. An uncontrolled trolley is rushing along the rails. Five hostages who cannot free themselves are tied to the tracks. If the trolley passes over them, all five will die. However, you can switch the arrow, and then the trolley will go to the sidings, which only one person is tied to. He is going to die, too. Will you switch the arrow knowing that by sacrificing one person, you will save five?

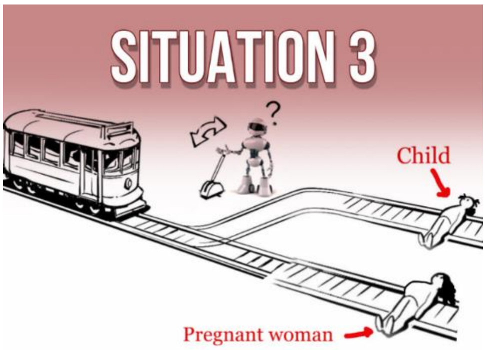


In this situation, most people will choose to kill one person and save the other five people. Robots are programmed for behavior aimed at the least human losses. There are robots capable of the simplest logic and able to "anticipate" possible outcomes of events and make the right choice corresponding to their algorithms. Thereby, the robot will also choose to save five people.

If we change the situation from the point of view of a person, and instead of one unknown person, a close relative of the subject will be on the tracks. A person under the influence of feelings can make the wrong choice from the point of view of morality and kill other people, saving the life of his relative. Robots do not have feelings, so this situation for them will be similar to the previous one. Consequently, the absence of feelings helps the robot to make the right moral choice.

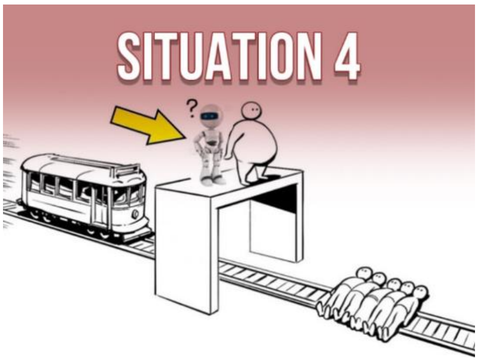


Let us change the situation again and put one person on each side: a child and a pregnant woman. According to the rules, the child's life should be a priority, so following the simplest logic, the robot will choose to kill an adult and it will be the wrong choice. A person, having the ability to think



broadly and possibly under the pressure of pity for an unborn child, can choose to kill one child and leave the other two alive.

Let us consider the fourth situation. The robot is standing on the bridge

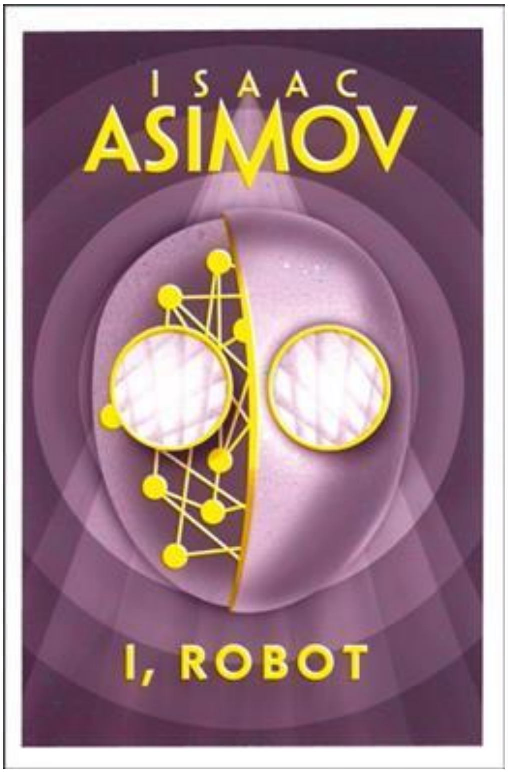


and sees that a trolley is moving along the rails under the bridge directly at five people who are likely to die now. Next to the robot there is a huge man who can be dropped in front of the trolley. The person will stop the trolley and save five people, but he will die. In this situation, a robot can turn from an observer into both a hero and a killer at the same time, but it is programmed according to the first law of robotics and cannot harm a person intentionally. Therefore, the robot will remain an observer and allow people to die. In addition, according to the first law, it can be accused of inaction, but it is actually in an unsolvable situation and has no choice. The robot has no freedom of action to commit premeditated murder, but also no responsibility because it is not a human and acts according to a clear algorithm.

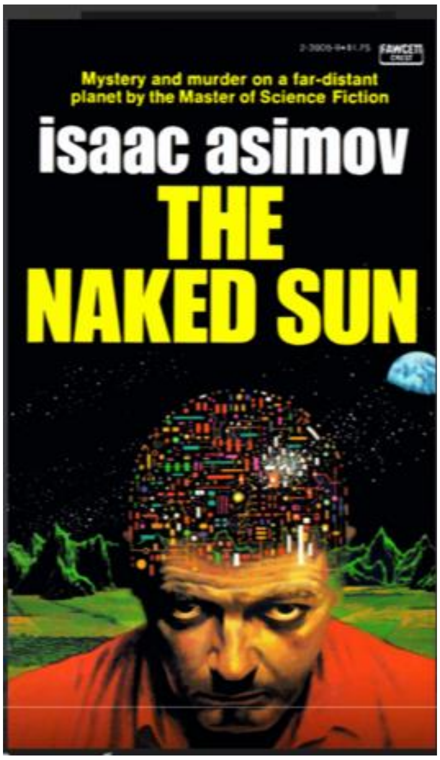
A real-world example might be Tesla's new autopilot mode. It works almost without errors under normal conditions, but in unforeseen situations, control is transferred to the driver. Thereby, the AI does not have algorithms for acting in case of all unforeseen situations and transfers responsibility to the pilot.

In conclusion, having studied all possible models of robots’ behaviour in difficult situations from the point of view of morality, the answer to the main question of the topic: where the limit between robot ethics and moral dilemmas is, can be given. Robots do not have feelings that are an important component of morality. Robots are unable to have full-fledged logic and act without algorithms programmed in them. The lack of free will does not allow robots to make full-fledged decisions in difficult situations. All these points are the line separating robot ethics and morality.

Appendixes



The cover of the book “I, Robot” by I. Asimov



The cover of the book “The Naked Sun” by I. Asimov

Bibliioghraphy

1. Issac Asimov. Evidence, 1946
2. Isaac Asimov. Foundation, 2022
3. Isaac Asimov. I, robot, 2018
4. Isaac Asimov. Runaround, 1942
5. Isaac Asimov. The naked Sun, 1978
6. https://www.tesla.com/support/autopilot
7. https://www.storyboardthat.com/articles/e/dilemma
8. https://www.frontiersin.org/articles/10.3389/frobt.2021.744590/full
9. https://www.researchgate.net/publication/314130573_Roboethics_-_Making_Sense_of_Ethical_Conundrums/fulltext/58b6cac1a6fdcc2d14d6c671/Roboethics-Making-Sense-of-Ethical-Conundrums.pdf