IA-I Report

on

The Trolley Problem and its Implications in Robotics

Submitted for course

Fundamentals of Robotics
(Minor in Robotics & AI)

By

Aryaman Gandhi

Roll No: 16010120185

Guide

Prof. Abhijeet Karmarkar

K. J. Somaiya College of Engineering
Vidyavihar, Mumabi - 400 077

2021-22

1. Introduction

1.1 What is the Trolley Problem?

The trolley problem is a famous thought experiment in ethics. It challenges one with a moral dilemma. The most common version of the trolley problem is as follows.

"A trolley (a train or tram) is moving towards five people on the main track. You are standing at a switch. If you turn the switch, the trolley will be diverted to a side track, but there is one person on this side track. Turning the switch will result in that person's death, and the five people on the main track will be saved. Not turning the switch will result in the death of the five people. Should you turn the switch?"

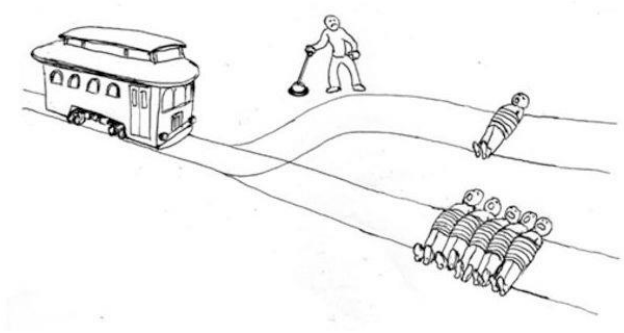

Fig (1.1) - An illustration of the trolley problem [7]

This problem has had several variations, all of which change the details of the question in one way or another. These modifications to the problem usually complicate the chain of causation and moral responsibility, and the opinions on the ethics of each of these situations depends on these details.

1.2 Importance of the Trolley Problem in Robotics

With the advent of modern technology and robotics, this seemingly implausible moral dilemma and the discussion surrounding it becomes much more important. One of the places where the

trolley problem is directly applicable is self-driving cars. It is easy to imagine how a situation similar to the trolley problem may arise in case of a self-driving car. Say the car is heading towards 5 pedestrians at a speed high enough where it is not possible for the car to come to a halt without crashing into the pedestrians. To protect them, the car may swerve onto the pavement, but in doing this it kills one pedestrian walking there. Alternatively, the car may crash into a tree or a building on the opposite side of the road, but that would risk killing the passenger of the car. Which of these three options would be the most "morally correct" option? In such a scenario, the designer of the algorithm that controls the self-driving car is in the same position as the person standing by the switch in the original trolley problem.

A deeper understanding of the trolley problem and the biases surrounding it allows the designers of the algorithms to make informed and ethical decisions when faced with such challenging moral problems.

2. Literature Survey

2.1 Previous Research

The trolley problem was first described by Philippa Foot in her 1967 paper, titled "The Problem of Abortion and the Doctrine of the Double Effect". [1] The doctrine of double effect states that if doing something morally good has a morally bad side-effect, it is ok to do it provided that the side effect was not intended. Many doctors use this doctrine to justify the use of high doses of drugs such as morphine for the purpose of relieving suffering in terminally-ill patients even though they know the drugs are likely to cause the patient to die sooner. In this paper, Mrs. Foot discusses abortion in relation to the doctrine of double effect, and the trolley problem is discussed introduced as an example. However, the version of the trolley problem described by Foot is different from the more common variant that was described above. In Foot's version, the reader is put into the seat of the trolley driver, rather than a bystander.

The "Bystander at the Switch" version of the trolley problem was conceived by Judith J. Thomson, in her paper titled "The Trolley Problem". [2] Thomson emphasizes that there is a difference between the two dilemmas. In Foot's version, the trolley driver is paid by the trolley company, and he bears responsibility for the safety of his passengers and anyone else who might be harmed by the trolley he drives. The bystander at the switch, on the other hand, is a person who just happens

to be there. According to her, the trolley driver would actively be killing the five people on the track by not taking action, while the bystander at the switch would not be killing anyone.

She also introduces the "Fat Man" variation. In this variation of the problem, the reader is standing on a footbridge above the trolley track. The trolley is moving towards the five people. On the footbridge, there is also a fat man, leaning over the railing. The reader has the choice to push the fat man off the bridge, in the path of the trolley. Doing so would certainly stop the trolley, thus saving the five people, but it would kill the fat man. Is it morally correct to push this man to his death to save five other people? According to Thomson, it is morally wrong to push the man as it infringes his rights.

The trolley problem is a topic of great discussion in the realm of self-driving cars, where the problem seems almost directly analogous to the case of unavoidable collisions in self-driving cars. However, a lot of literature warns against making this tempting analogy. In [3], the authors explain in detail the differences between the "split second" decision that must be made by the bystander in case of the trolley problem, and the contingency planning done by the programmer of the self-driving car. The case of the trolley problem is idealized and specific, but in real life such clear cut situations never arise. In real-life scenarios, the relationships between the different elements of the problem tends to be a lot more complex and there are many more variables involved, and the final decision stems from a mix of risk estimation and decision making under uncertainty. Additionally, while the Trolley Problem only deals with the aspect of moral responsibility, in real life one must bear moral as well as legal responsibility. Regardless of this, it is still worthwhile to entertain the Trolley Problem as a stepping stone or a "starting point" for the development of ethical self-driving cars. In [4], the authors lay down a mathematical framework combining Bayesian, Equality and Maximin principles. This mathematical model may serve as a final "solution" to the trolley problem.

2.2 Experiments

Researchers placed a number of test subjects in a 3D simulated virtual reality recreation of the trolley problem. [5] Participants stood in a dimly lit, sound-resistant room, where they wore a head-mounted display device that transmitted video and audio directly to the eyes and ears. The virtual environment unfolded with the participant standing on a platform overhanging a railway track. Behind the participant, a main track stretched to the horizon, while in front, the main track

split into two tracks, one that continued straight through a ravine, and a side track that veered off through another ravine. Directly in front of the participant was a rail switch, manipulated via a force-feedback joystick. The procedure began with several trials in order to habituate participants with the environment and task, followed by an experimental trial, and then three exploratory, post experimental trials. The outcome of the experiment was that about 90.5% of the participants chose to pull the switch. [5]

3. Observations & Discussion

From the experiment described in 2.2, it is evident that most of the participants, and by extension most people, fall under the "utilitarian" camp. The utilitarian perspective is that of numbers, it values 5 lives more than 1. This utilitarian worldview is supported by the mathematical framework described in [4] too.

However, I feel that the issue of the trolley problem in relation to self-driving cars is way more complicated and nuanced than as described by Thomson. This is well discussed in [3]. It is important to remember that the conditions in the trolley problem are *ideal*. This implies that everything is certain, the only variability that can occur is the choice of the bystander at the switch. In real life however, things are not clear cut. The AI system of the self-driving car does not have all the information, and it does not know how reliable the information gathered from the sensors is. For example, in real life a situation may occur where the AI is only 60% confident that it will kill someone if it turns. Focusing on the self-driving car itself, in order to calculate the optimal trajectory, the self-driving car needs (among other things) to have perfect knowledge of the state of the road. But even very good data from advanced sensors can only yield estimates of the road's exact condition. Moreover, regarding each of the five workers: their chances of surviving the collision with the car depends on many factors, for example their age, their overall state of health. The car's technology might enable it to gather partial information, but it will never know all the variables.

It is also very important to acknowledge the difference in the position of the "effector" in each case. As stated earlier, in the trolley problem the bystander is just a person who happened to be

present next to the switch, while the algorithm designer is a person consciously deciding the outcome of a trolley-problem like situation, were it ever to occur. The algorithm designer has access to sufficient time and knowledge to make an informed, ethical decision. Also, realistically, the decision-making about self-driving cars is more realistically represented as being made by multiple stakeholders – for example, ordinary citizens, lawyers, ethicists, engineers, risk-assessment experts, car-manufacturers, etc. These stakeholders need to negotiate a mutually agreed-upon solution. To illustrate the bearing this has, let us revisit the hypothetical situation that was described in 1.2.

A self - driving car is heading towards 5 pedestrians at a speed high enough where it is not possible for the car to come to a halt without crashing into the pedestrians. To protect them, the car may swerve onto the pavement, but in doing this it kills one pedestrian walking there. Alternatively, the car may crash into a tree or a building on the opposite side of the road, but that would risk killing the passenger of the car.

If a person was driving the car, then possibly the most morally upright decision would be that of self-sacrifice. However, this does not carry over to a self-driving car. It would be a self-driving car that intentionally jeopardizes the safety of the passenger, and nobody would buy it.

Another facet highlighted in [3] is the legal responsibility of a self-driving car. With the occurrence of serious crashes and collisions – especially if they involve fatalities or serious injuries – people are disposed to want to find some person or persons who can be held responsible, both morally and legally. This is an unavoidable aspect of human interaction. If a self-driving car is to hit someone and kill them, who would be legally responsible? Would it be the designer of the algorithm? Or would it be the corporation that manufactures the cars? Would the passenger be held accountable? Or would the AI itself be the one to blame for this death? This is a completely separate question that deserves its own paper, but the point being made is that this is an entirely different aspect of the trolley problem that is completely ignored in the general discussion of it.

Also, I think that it is also important to know the nuances of how the self-driving car is actually programmed. Most self-driving car AI systems use a Partially Observed Markov Decision Process (POMDP). [6] The decisions are made by the car with the help of the data being fed in through the

sensors, along with the past data. A common way to train an AI is with the help of something called reinforcement learning. In reinforcement learning; an AI is given a task to do. If the task is performed correctly, it is given a "reward" (positive reinforcement) or a "cost" (negative reinforcement). The AI tries to better itself by moving in the direction that offers an increased reward or a reduced cost. It is the reward or cost function that ultimately determines what the AI learns. If a function is designed to give the AI a high reward for performing task X, then the AI will eventually learn to do task X. So, in the case of the trolley problem, the AI would first have to be trained for it to learn how to react to it.

So, the AI would be put through many iterations of a simulated version of the trolley problem. If it makes a decision that is appropriate according to the person training it, then the AI is rewarded. Eventually, this AI will learn to react to the trolley problem in the way it has been trained to. And then, if it encounters such a problem in real life, it will act accordingly. The important takeaway here is that the AI is not really assigning an importance to the lives of any of the six workers involved in the problem, and neither is it making a conscious choice to save anyone. It is just repeating what it has been trained to do. As such, it is difficult to train an AI to be "moral" or "ethical", since the AI's version of morality would heavily derive from the morality of the person or group training the AI.

In conclusion, the trolley problem, while being a very interesting topic that is still relevant today, does not apply as cleanly to the concept of robotics or self-driving cars. However, the problem is still a good stepping stone or a foundation to the world of the ethics of self-driving cars, so it is still worthwhile to study it.

4. References

[1] Foot, P. (1967). The problem of abortion and the doctrine of the double effect. Oxford review, 5, 5–15. https://doi.org/10.1093/0199252866.003.0002

[2] Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 94(6), 1395–1415. https://doi.org/10.2307/796133

[3] Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

[4] Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk. *Philosophy & Technology*, *34*(4), 1033–1055. https://doi.org/10.1007/s13347-021-00449-4

[5] Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: emotion and action in a simulated three-dimensional "trolley problem". Emotion (Washington, D.C.), 12(2), 364–370. https://doi.org/10.1037/a0025561

[6] Roff, H. M. (2022, March 9). *The folly of trolleys: Ethical challenges and autonomous vehicles*. Brookings. Retrieved April 6, 2022, from https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/

[7] Feldman, B. (2016, August 9). *The Trolley Problem Is the Internet's Most Philosophical Meme*. Intelligencer. Retrieved April 6, 2022, from https://nymag.com/intelligencer/2016/08/trolley-problem-meme-tumblr-philosophy.html