Proceedings of the 4th International Conference on Educational Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/15/20231039

Autonomous Driving Ethics: Self-driving Cars Facing Trolley Problems

Yicheng Xiang[1,a,*]

[1]*Faculty of Law, East China University of Political Science and Law, Guangfulin Street, Shanghai, 201620, China*
a. 212125010453@ecupl.edu.cn
**corresponding author*

Abstract: The trolley problem has always been one of the most famous ethical dilemmas in human history, arousing millions of arguments and discussions. When self-driving cars, as products of our new age, encounter the trolley problem, the decisions made will be different from those of human beings due to the particularity of the cars and the autopilot algorithm. This paper will first determine the definition and scope of the trolley problem, and then compare the difference between autonomous cars and traditional ones. It then proposes possible guidance for the autopilot algorithms from legal and ethical perspectives. Despite a long history, ethical methods for autonomous cars are still uncertain and hard to be accepted by everyone. Perhaps the combination of different ethical approaches could greatly ameliorate this problem. The feature of hysteresis of law makes it impossible to provide guidance for autonomous cars in time, but laws can help build abstract principles for autopilot algorithms from the perspective of reducing legal liability.

Keywords: trolley problem, autonomous driving, autopilot algorithm, ethical approaches, legal liability

1. Introduction

Problems relevant to law and morality have been the core arguments of legal studies. Among the related topics, the trolley problem is one of the most well-known virtual cases and has aroused lots of discussions between public and the academic circles. Each point of view depends on their own legal and moral rationales. Thus, whether the party in the case is innocent or guilty and whether he or she should be punished or not has so far been inconclusive. But when it comes into practice, those theoretical problems transform into numbers of living lives, which requires the decision makers to choose the better option.

In this day and age, self-driving cars are increasingly widespread, accompanied by cutting edge technology and algorithm as well as hidden danger and accidents. There is no denying that the artificial intelligence technology (AI) is advanced enough to work out most of the potential cases, but when it comes to the moral dilemmas, such as the trolley problem, completely rational thinking of the AI will make a big difference compared with humans.

The trolley problem is by nature a conflict between legal interests, while self-driving cars and the AI are related to conflicts between individual autonomy and automated decisions. As a result, under the circumstance where a self-driving car is faced with a trolley problem, a set of algorithms

DOI: 10.54254/2753-7048/15/20231039

formulated by human must be applied to ensure the minimum loss. Obviously, it has been a puzzle for countries, societies, and those algorithm makers. Current research stops with an abstract principle which only requires technology for protecting human rights while lacking specific solutions for moral dilemmas and conflicts of interest. At the same time, due to the feature of hysteresis of law, regulations and laws are unable to provide a timely guidance for algorithm.

In hence, a significant number of theories and attempts are put forward at the aim of measuring legal rights and make the better option of who to hit. The paper will discuss the trolley problem in the case of automatic driving from the perspective of laws and morals. The purpose of the paper is to analyze and conclude a better guidance for autopilot algorithm.

2. Legal and Ethical Guidance

2.1. The Trolley Problem

The trolley problem originated from a circumstance irrelevant to autonomous driving. Whether a train driver should or should not change the orbit to save the five workers on track by killing the one on the other track was the pioneering virtual case raised by the British philosopher Philippa Foot. Another philosopher Judith Jarvis Thomson attempted to compare the virtual case with the following real legal case. A surgeon called Rodney Mapes deliberately killed an innocent person with the aim of respectively transplanting the five organs to five dying patients, which triggered a mass of disgust and objections. Eventually, he was accused of murder [1].

The reason why the two cases look alike is that the driver and the surgeon both killed one and saved five. However, both options in the first case are seemingly acceptable, while the surgeon's behavior is regarded as a murder against humanity. The internal difference is that the source of danger. Only if the source of danger is sole and concurrent, the case can be defined as a trolley problem. In the case of the surgeon, the risk of the five patients is not brought by the surgeon. Furthermore, if the innocent one was not killed, the five would not die at the same time. Thus, the trolley problem mentioned in this paper is scoped [2].

From a legal and moral perspective, the trolley problem is interpreted in two opposite views. In 2012, a woman called Jones witnessed a trolley that is out of control. She hit the switch and changed the direction of the trolley, costing a person's life to save five workers' lives. Those who insist Jones was guilty suggest that each man is born equal, and no one has the right to determine the lives of others. And it should be rejected to protect the majority's profit and happiness at the cost of the minority, or it can be an excuse for violence and crimes. The opposite point proposes that from a utilitarian view, Mrs. Jones's behavior is permitted since she guaranteed the maximization of happiness and profit. In addition, according to Summa Theologica written by Thomas Aquinas, under certain circumstances, one's behavior is acceptable if it is ethical but accompanied by a bad result [3]. In this case, Mrs. Jones's behavior was well-meaning, and she did not subjectively pursue the person's death. Thereby, her behavior can be considered ethical and innocent.

2.2. The Particularity of Autonomous Driving

Obviously, the virtual case of the trolley problem totally differs from the real driving scene, which is reflected in the time for collision decision-making. In traditional driving situations, on one hand, human drivers can respond flexibly, such as braking and changing direction at the same time to avoid accidents in advance to the greatest extent. On the other hand, most of the decisions made by human drivers come from spontaneous and instantaneous instinctive reactions. As long as the damage is not intentionally expanded, and whether their aim is to protect others or themselves, illegality can be excluded. In contrast, the decision made by AI is a result of pre-design and

DOI: 10.54254/2753-7048/15/20231039

contingency plan of the algorithm. Furthermore, due to its feature, the AI's decision-making is not a single task, but a process that needs to consider the needs and demands of all parties, which will directly or indirectly affect each party relevant to the accident.

Apparently, the best way to solve the decision-making problem of the automatic driving is to prevent the moral dilemma from taking place. Despite an increasingly improved algorithm which reduce the risk of collision to a great extent, the probability of making ethical decision still exist, regardless of strong precautionary measures. Due to the infinitely growing marginal cost, it is impractical to eliminate the risk of dilemma completely, so methods should be taken to apply a better guidance for autopilot algorithm, both from law and ethic.

For the purpose of this paper, the vehicles are assumed to be fully automated. Therefore, there is no consideration about the further issue of whether, when or how the vehicle should pass control back to the human driver. Furthermore, there will be no human driver in the automated vehicles. Anyone who is in the vehicle will be considered a passenger, without any responsibilities for the car's behavior [4].

2.3. The Function of Law

Germany already has a relatively mature set of ethical principles for autonomous driving. Firstly, §1b of the Eighth Amendment to the Road Traffic Act of 2017 sets out the liability and obligations of drivers in autonomous vehicles:

(1) The driver of the vehicle may no longer concentrate while driving the vehicle using the highly or fully automated driving function in accordance with paragraph 1a in traffic conditions and vehicle control; In doing so, he must remain sufficiently vigilant so that he may discharge his obligations under subparagraph (2) at any time.

(2) The driver of the vehicle is obliged to immediately regain control of the vehicle in the following cases:

a. if the highly or fully automated system requests him to do so, or

b. if he realizes, or due to obvious circumstances, that the prerequisites for the intended use of the highly or fully automated driving function are no longer met.

At the same time, it also clarifies the allocation of responsibility for autonomous vehicle accidents:

If a traffic accident occurs while using the automated driving function under §1b(1), the autonomous vehicle manufacturer shall be liable for compensation if the liability lies with the autonomous vehicle; If, in the case of §1b(2), the driver should have taken over but failed to take over, or if an accident occurred after the takeover and the responsibility lies on the autonomous vehicle side, the driver shall be liable.

In step with the Amendment, Germany has released the regulation of the ethics for automatic driving. In 2016, Germany established a committee called Ethics Commission on Automated and Connected Driving in order to provide guidance on ethical issues in driverless and network-connected driving systems. One year after it was established, 20 ethical principles were published, becoming a reference for autonomous driving industry all over the world. While reaffirming human autonomy, these principles have several matters of concern:

One that is less controversial is the question of value ordering, with human life taking precedence over the safety of animals and property. When one of the two humans and animals has to encounter a danger to their lives, priority is given to protecting human life.

But when AI is faced with the human-to-human trolley problem, the situation becomes complicated. The eight ethical principle acknowledges the non-standardizable nature of autonomous driving decisions, namely that "True ethical dilemma decisions, such as life-to-life trade-offs, depend on specific factual contexts that include the unpredictable behavior of all participants. Such

DOI: 10.54254/2753-7048/15/20231039

decisions cannot therefore be standardized and programmed according to an unambiguous ethical system." That is, the law allows self-driving car drivers to be distracted and no longer focused on driving after the AI takes over driving, but when ethical dilemmas arise, such as either not turning right and hitting the three students crossing the street or turning right and hitting the delivery man riding an electric car, the AI will send a takeover request to the driver. At this point, requiring an inattentive driver to make a judgment about an unexpected situation in a sudden is likely to maximize the loss.

The defect above forced the commission to propose new principles that recognized the limits of the drivers faced with the ethical dilemmas and demanded that "the software and technology of highly automated vehicles should actually avoid the concretize risk allocation policies sudden transfer of control to human drivers in an emergency. In order to ensure effective, reliable and safe human-machine communication and avoid overburdening people, systems should be more adaptable to human communication behavior and not require high human adaptability." The author believes that this principle contradicts with the eighth principle. Admittedly, it is not advisable for autonomous driving systems to hand over driving rights to the driver in an emergency. But the eighth principle has already acknowledged the limit of autonomous driving systems that they cannot substitute for human ethical decision-making, and they do not have the right to weigh the value of human life. Summarizing these two principles, it can be found that neither the decision can be handed over to the driver, nor can the decision be made automatically by the autonomous driving system. If machines can foresee the dilemma and ask drivers to take over with plenty of time to prepare, it cannot be called as a true ethical dilemma, or even a real ordinary traffic accident. Most accidents take place in a sudden without any preparation [5].

From a legal and principled point of view, human taking over is naturally the best option, but this requirement clearly overburdens the driver and does not meet the real situation of the trolley problem.

2.4. Ethical Approaches

Therefore, in the decision-making of the inevitable accidents of autonomous driving, due to its hysteresis, the law cannot realistically guide autonomous vehicles in advance. Thus, ethical approaches are needed. Laws and ethics play different roles in the society. It is obviously unreasonable for the law and regulations to require those impossible, or it will be an imposition for citizens. But it is permitted that the ethical and moral principles can put forward a higher demand. The ethic is something like a teacher making impossible demands for his students with the aim of inspiring their potential. Therefore, the author believes that laws and principled regulations cannot effectively solve this trolley problem, which ethical elements and theories must be added to complement legal rules to solve this problem better. In this background, there are several ethical approaches:

2.4.1. The Deontological Approaches

Firstly, the deontological approaches. Deontological ethics determines the right ethical behavior by constructing a set of rules, and its key feature is that rules can be layered by setting clear priorities. Isaac Asimov's Three Laws of Robotics is a classic example of deontological ethics, stating that:

First Law: A robot may not injure a human being, or, through inaction, allow a human being to come to harm. Second Law: A robot must obey orders given it by human beings, except where such orders would conflict with the First Law. Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law [6].

DOI: 10.54254/2753-7048/15/20231039

The theory of deontological ethics provides a structure which plays the role as constraints in decision-making and algorithms, such as restrictions that prohibit vehicles from actively harming humans and property, prohibiting violations of specific traffic laws, and so on. Collision algorithms that apply deontological principles can guarantee that self-driving cars will operate according to certain rules, and while there are certain rigidities that potentially and inevitably cause harm, they also make the car's behavior in extreme situations predictable and explainable.

However, at the technical level, the application of deontology rules has great limitations, rules such as "cannot harm humans" in the Three Laws of Robotics are too idealistic, and in real driving scenarios, complex factors should be considered comprehensively. Thus, if deontological principle is the sole rule to follow, the needs and the safety of all the parties may not be guaranteed. In addition, At the level of laws and regulations, the deontological principle is usually closely linked with the current traffic laws and regulations, and overcoming the ambiguity of legal text expression and making it clearly expressed in machine language in algorithms is a great challenge for collision programming [7].

2.4.2. The Utilitarian Approaches

Secondly, the Utilitarian approaches. A set of cost function algorithms will be designed to calculate the expected costs for various possible options and select the one that involves the lowest cost. For instance, to sacrifice the minor number of victims in car crashes. Seemingly, it is more effective and less controversial. However, it depends on the exact definition of the benefit being maximized, which varies immensely from country to country. Furthermore, as an optimization problem that minimizes collective rather than individual harm, the extreme case algorithm may use inappropriate features as decision-making criteria in actual operation, resulting in biased risk allocation decisions [8].

2.4.3. Summary

Therefore, both approaches have limitations. If the principles proposed by the deontological theories are combined with the choices supported by the utilitarian theories in a reasonable way, they may produce the effect which may be greater than that of their own. For instance, to take the option which involves the lowest cost under the premise of following the deontological principles.

2.5. Moral Logic in Autopilot Algorithm

Under such circumstances, scholars have attempted to explore and conclude public's moral preferences and apply those moral choices into the autopilot algorithm. In other words, to make AI construct their own rules and values as well as find a better solution by learning from human's judgment and driving experience. AI's studying from a large scope of data systematically enables it to imitate human drivers in extreme cases like trolley problems. As a result, the decision-making will potentially be more acceptable and ethical.

2.5.1. The Moral Machine Experiment

Professor Edmond Awad at the University of Exeter once proposed a large-scale ethical experiment called the Moral Machine for autonomous driving scenarios, which collected approximately 40 million questionnaires about the collision option they will select when facing a trolley problem. The experiment covered from 233 countries and regions, including more than 10 types of languages. Among the 9 moral choices, most people prefer to protect human rather than animals, the young rather than the elder, the majority rather than the minority. But to some of the other moral choices,

DOI: 10.54254/2753-7048/15/20231039

such as whether healthier ones are prior to those who have underlying diseases and whether those of highly ranked social status are prior to those of lower-ranked status, the answers varied hugely due to the difference between respondents' living places, cultures, financial means and educational attainments [9]. Thus, the result of the experiment is unable to be applied in the algorithm, or the algorithm will be influenced significantly by the quality and quantity of the respondents, and no one will get satisfied with the autopilot algorithm. Furthermore, setting the moral preference for AI artificially violate the current moral principles such as Article 9 of the Ethical Guidelines for Automated and Connected Driving by the German Ethics Commission on Automated Driving, which presented that in the event of unavoidable accident situations, any qualification based on personal characteristics (age, gender, physical or mental constitution) is strictly prohibited.

2.5.2. Embedding the Driving Data of a Human Driver

Since the Moral Machine Experiment is not perfect and fair enough to balance the options from people all over the world, it seems that another way to solve this problem is to teach AI to drive and think like a real person by embedding the driving data of a human driver to the AI. Even though it can eliminate the defect of the Moral Machine Experiment, AI is likely to imitate those unethical behaviors from barbarous and furious drivers, which is worthy of caution, especially in a life-threatening situation [10].

2.6. Abstract Principles

Some abstract principles and concepts can be preliminarily put forward as guidance. These abstract principles should not only be universal and easily accepted by the public, but more importantly, reduce the severity of the consequences and the legal liability.

Once self-driving vehicles involve criminal responsibility, their manufacturers and algorithm designers are the bearers of the responsibility due to the special legal status of autonomous vehicles and AIs. In the autonomous driving scenario, where there is no longer elimination of misfeasance, algorithm designers should attempt to avoid risks as well as legal liability.

2.6.1. Intentional Homicide

In order to avoid the decision of the algorithm from constituting the crime of intentional homicide, the first principle is that an innocent third party cannot be sacrificed with the aim of protecting the driver himself. In traditional driving scenarios, when the driver sacrifices a third person's life to save himself, it cannot be recognized as a legitimate emergency hedging. The justification of emergency hedging is based on social joint obligations. Due to the joint between the third party and the member who encounter danger, the emergency hedging should be legitimated to a certain extent [11]. For the autopilot algorithm which can be designed in advance, legal judgment will be more stringent [12]. In the autonomous driving scenario, for an innocent third party, sacrificing his life and legal interest is obviously beyond the scope of social joint obligations. Thus, emergency hedging cannot be justified. Therefore, manufacturers must regard the principle of avoiding sacrificing innocent road traffic users as one of the priority rules when designing autonomous algorithms.

2.6.2. Traffic Offense

In the stage of constructing the algorithm for extreme cases, manufacturers must exclude the possibility of a traffic offence as much as possible. According to the Criminal Law, the crime of traffic offence refers to the violation of traffic and transportation management regulations, resulting

DOI: 10.54254/2753-7048/15/20231039

in a serious accident and leading to either serious injury and death or public and private property losses. Violation of traffic and transportation laws, as one of the constituent elements of traffic offense needs to be excluded, requiring self-driving cars to strictly abide by traffic laws when dealing with extreme situations. In addition, under the premise of not constituting the crime of intentional homicide, if the self-driving car inevitably violates traffic rules and causes traffic accidents, it should choose a plan that causes a lower number of deaths and serious injuries to mitigate the possibility of constituting a traffic offense [7].

3. Conclusions

The scene where autonomous driving and the trolley problem meet is not only relevant to personal safety, but also a dilemma which is extremely difficult to be solved perfectly. In current situations, the core of the different principles and solutions is to reduce risks and damages. The chances of survival are unevenly distributed before the system selection is decided. The autonomous driving dilemma is actually a statistical distribution of risks, and it is necessary to combine relatively abstract principles to gradually concretize risk allocation policies. Therefore, the combination of established ethical approaches and abstract principles gradually explores more reasonable risk allocation schemes, which provides feasible ideas for the ethical problems of autonomous driving even if there are defects.

References

[1] *Andrade G. (2019) Medical ethics and the trolley Problem. J Med Ethics Hist Med. Mar 17; 12:3. PMID: 31346396; PMCID: PMC6642460.*

[2] *Bruers, S., Braeckman, J. (2014) A Review and Systematization of the Trolley Problem. Philosophia 42, 251–269.*

[3] *Thomas Aquinas. (2013) Summa Theologica. Shanghai, The Commercial Press.*

[4] *Lawlor, R. (2022) The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases. Res Publica 28, 193–216.*

[5] *Zheng Ge. (2022) Trolley Problem and Ethical Considerations in Algorithm Design for Automated Vehicles. Zhejiang Social Sciences, No.316(12):37-47+67+156.*

[6] *Anderson, S.L. (2008) Asimov's "three laws of robotics" and machine metaethics. AI & Soc 22, 477–493.*

[7] *Wei Xinquan. (2023) Research on algorithm governance in extreme situations of autonomous driving. Cybersecurity and Data Governance, 42(03):38-45.*

[8] *Geisslinger, M., Poszler, F., Betz, J. et al, (2021) Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk. Philos. Technol. 34, 1033–1055.*

[9] *Awad, E., Dsouza, S., Kim, R. et al. (2018) The Moral Machine experiment. Nature 563, 59–64.*

[10] *Lim H， Taeihagh A. (2019) Algorithmic decision-making in AVs: understanding ethical and technical concerns for smart cities. Sustainability, 11(20): 5791.*

[11] *Chu Chencheng. (2018) The Criminal Legitimacy of Solving the "Trolley Dilemma" in Automatic Car Programming. Global Law Review, 40(03):82-99.*

[12] *Wang Gang. (2011) The Duty of Tolerance and Limits of the Innocent Third Party in Emergency Avoidance and the Justification of Emergency Avoidance. Peking University Law Journal, 23(03):609-625.*