World Electric Vehicle Journal

Article

Ethical Considerations of the Trolley Problem in Autonomous Driving: A Philosophical and Technological Analysis

Hao Zhan [1,*] **and Dan Wan** [2]

[1] Department of Philosophy, Xiamen University, Xiamen 361005, China
[2] Chinese Ethical Civilization Research Center of Hunan Normal University, Changsha 410006, China; dan.w@hunnu.edu.cn
* Correspondence: haozhan1993@gmail.com; Tel.: +86-177-3801-7758

Abstract: The trolley problem has long posed a complex ethical challenge in the field of autonomous driving technology. By constructing a general trolley problem model, this paper demonstrates that the default loss assumption is a necessary condition for the occurrence of trolley problems. However, an analysis of the differences between classical trolley problems and autonomous driving scenarios reveals that this assumption is not supported in the design of autonomous driving systems. This paper first provides a detailed definition of the trolley problem within the context of autonomous driving technology and establishes a general trolley problem model to better analyze the issue. We then discuss two solutions: the first solution acknowledges the existence of the trolley problem in the context of autonomous driving technology but does not recognize the existence of a "most acceptable decision"; the second solution denies that decision-makers are limited to a finite number of decisions, each resulting in a corresponding loss. Based on the second solution, we propose a "sufficient time" solution, illustrating that the interaction between planning and control systems in autonomous driving can avoid ethical dilemmas similar to the trolley problem. Finally, we analyze from a philosophical perspective why the trolley problem does not arise in the context of autonomous driving technology and discuss the ethical responsibilities associated with autonomous driving. The design goal of autonomous driving technology should be a zero-accident rate, which contradicts the unavoidable loss assumption in the traditional trolley problem. Therefore, the existence of the trolley problem is unrealistic in the practical application of autonomous driving technology.

Keywords: autonomous driving technology; autopilot; trolley problem; ethics responsibility

check for updates

Citation: Zhan, H.; Wan, D. Ethical Considerations of the Trolley Problem in Autonomous Driving: A Philosophical and Technological Analysis. *World Electr. Veh. J.* **2024**, *15*, 404. https://doi.org/10.3390/wevj15090404

Academic Editor: Ayman EL-Refaie

Received: 12 July 2024
Revised: 7 August 2024
Accepted: 22 August 2024
Published: 4 September 2024

1. Introduction

In recent years, autonomous driving has attracted attention worldwide. Traditional manufacturers, such as Audi and Volvo, as well as Internet companies, such as Uber, Baidu, and Waymo, have all started to become involved in the development of autonomous driving.

On the one hand, with the continuous breakthrough of autonomous driving technology (ADT), the mass production of self-driving cars is no longer a fantasy. On the other hand, countries have begun to legalize ADT through legislation [1]. These signs seem to suggest that the complete application of ADT is just around the corner.

However, with the popularity of ADT globally, the social science field began to debate it. The philosophical area raised the most characteristic controversy with the trolley problem in autonomous driving technology [2].

MIT launched the "Moral Machines" platform in 2016. People (nearly 40 million) from 233 countries were surveyed about their choices in the face of various kinds of trolley problems. It was found that there were some common consensuses. For example, respondents generally chose to protect human beings over other animals, chose to protect the lives of the many rather than the few, chose to protect the young rather than the old, and chose to protect those who obey traffic rules rather than those who violate them [3,4].

At the same time, scholars are also keen to explore solutions for self-driving vehicles facing the trolley problem through a "moral machine". They propose moral algorithms to solve those problems [5]. Some works discuss the social dilemmas posed by moral machines: as citizens, people hope that ethical algorithms can save more people according to utilitarianism; as consumers, people are more likely to buy cars that keep drivers and passengers in the vehicle safe [6].

All these problems make us wonder whether the trolley problem in the context of ADT can be solved. Is the trolley problem in the context of ADT a real problem?

ADT driven primarily by AI systems represents a new mode of driving and raises novel issues of legal liability. Among the most significant is the difficulty in determining the responsible party in the event of a traffic accident and whether the autonomous driving system can be held accountable as a liable party.

Marchant and Lindor argue that in autonomous driving scenarios, whether partial or fully autonomous, the allocation of responsibility between the autonomous system and the driver is usually exceedingly difficult [7]. They analyzed traditional accident scenarios, noting that common causes of traffic accidents are mainly due to driver error, vehicle malfunction, and unavoidable natural circumstances, with the driver and car manufacturer typically being the potential liable parties. In traditional scenarios, assigning responsibility is relatively straightforward. However, in autonomous driving scenarios, regardless of whether it is partial or full autonomy, the responsibility allocation between the autonomous system and the driver becomes exceedingly difficult.

Some studies emphasize the diversity of responsible parties. Douma and Palodichuk, as well as Si Xiao et al., have more broadly discussed the practical scenarios and consequences of various potential liable parties, including manufacturers, drivers, and insurance companies [8,9].

Some studies focus on the diversity of resolution strategies. They discuss several ethical theories that can be applied to the trolley problem in autonomous driving scenarios, including utilitarianism, deontology, or Kantianism [10]. Other research suggests that in such situations, random selection should be used, arguing that random selection, as an action without choice, represents a respect for life [11]. Further studies examine the real-world consequences, such as strict product liability, that may arise after autonomous vehicles make decisions in moral dilemmas [12].

It is worth noting, however, that in their studies on the issue of liability, researchers generally approach autonomous vehicles by directly transplanting AI systems into normal cars without paying sufficient attention to the unique context of autonomous vehicles. This approach often results in using ethical choices for AI systems to judge the liability issues of autonomous vehicles. However, the liability issues of autonomous vehicles are different from general AI issues, involving more complex, specific circumstances. Therefore, we need to rethink and specifically analyze the liability issues raised by autonomous vehicles.

In this work, we first analyze the specific definition of the trolley problem in ADT.

Second, we build a general trolley problem model to analyze the trolley problem better and determine the core factors leading to the dilemma.

Third, we try to provide three different ways to solve this problem.

Fourth, we try to represent these core factors in the autopilot scenario. We find that the core presupposition leading to the dilemma is contradictory to automated driving technology. From this, we can conclude that the trolley problem will not occur in the autonomous driving technology scenario.

Finally, we further analyze, from a philosophical perspective, why the trolley problem will not appear in the context of autonomous driving technology.

2. Traditional Trolley Problem

The trolley problem as a thought experiment was first proposed by Philippa Foot [13]. Foot asked us to imagine that a driver faced only two choices in an uncontrollable tram: one choice was not to change the direction of the tram, which would cause it to run into five people; the other choice was to change the direction of the tram, which would sacrifice one person on the other track. The question was, "What is your choice?".

Judith Jarvis Thomson came up with some improved versions based on Foot, and the most famous one is the George trolley problem [14]:

George is on a footbridge over the trolley tracks. He is familiar with trolleys and can see that the one approaching the bridge is out of control. On the bridge's track, there are five people; the banks are so steep that they will not be able to get off the track in time. George knows that the only way to stop an out-of-control trolley is to drop a hefty weight into its path. However, the only available, sufficiently heavy weight is a man, also watching the trolley from the footbridge. George can shove the man onto the track in the trolley's path, killing the man, or he can refrain from doing this, letting the five people die.

This ethical topic abated after a period of heated discussion. Unexpectedly, it has been raised again, after the rise of ADT in recent years, as a criticism of emerging technologies and has aroused the public's and academia's attention. Compared with the classic topic of the relationship between positive duty and negative duty, people are increasingly concerned about the decisions that self-driving cars make when facing the trolley problem today. The consequences of these decisions will be effective not only in the philosophers' minds but also in the real world.

Jean-Francois Bonnefon compared various decision-making methods and proposed that ADT should have consistency (being consistent in every situation), stability (not causing public outrage), and feasibility (not discouraging buyers).

In a public survey, he found that the majority of non-professionals, following utilitarian views, believe autonomous cars should be diverted in the trolley problem, minimizing the number of casualties [5]. He also delved into various scenarios of the trolley problem in the context of ADT, including a situation involving the death of the decision-maker himself/herself and a situation involving senior citizens and children, in an attempt to come up with morally acceptable algorithms [6].

However, some researchers object to this conclusion based on public opinion and believe that ADT should build a Rawlsian algorithm based on Rawls's maximin principle and make decisions accordingly [15]. Other studies do not evaluate the specific choice made by autonomous cars facing the trolley problem but only give affirmation to the view that ADT has to make an autonomous judgment [16]. Some studies hold that embedding ethical principles into ADT and guiding self-driving cars in making decisions when facing the trolley dilemma is imminent [17].

It is difficult to judge which of the above studies people should accept. To solve this problem, we first need to review the essence of the trolley problem.

The unchanging essence of the trolley problem is called the "ethical dilemma", which remains consistent from Foot's trolley problem to Bonnefon's trolley problem in the context of ADT. It is not difficult to summarize the common characteristics of the trolley problem:

1. The limit-inevitability of selection. The options are limited, and a selection must be made.

2. The inevitability of loss. No decision can avoid loss.

3. The complexity of evaluation. It is not easy to judge which decision is more acceptable.

Someone may say the question in Foot's trolley case is not "What is your choice?" Recall that in Foot's paper, the trolley example is contrasted with another example called Transplant, in which a doctor has the option to kill one patient and harvest his or her organs to save five other patients. Foot's point is that while it seems permissible to save the five people in the trolley case, it seems impermissible to save the five patients in the transplant case. Hence, there must be a morally significant difference between the two cases. Foot uses the two cases to argue for a positive vs. negative duty distinction. In the trolley case, the trade-off is a positive duty to save one vs. a positive duty to save five, and the positive duty to save five outweighs the positive duty to save one. In contrast, in the transplant case, there is a negative duty not to kill one vs. a positive duty to save five. The negative duty, Foot argues, is sufficiently strong to outweigh the positive duty to save five.

However, in this paper, we are not discussing the traditional trolley problem per se but the moral dilemma abstracted from the trolley problem. The point of trolley cases in moral philosophy is not to examine what we ought to do should a trolley case in fact arise. The point is to consider which scenario features make a difference in the moral permissibility of the agent's actions under idealized circumstances. Thus, whether autonomous vehicles will encounter trolley cases in practice is mainly irrelevant to the question of whether trolley cases can inform the ethical design of autonomous vehicles. However, our work examines situations in which the trolley case in fact arises. That is the difference with moral philosophy.

3. General Trolley Problem Model

On the basis of the specific definition, we can build a general trolley problem model to analyze the trolley problem better and determine the core factors leading to the dilemma.

There is a driving scenario S in which the decision-maker can only execute a limited number of decisions, and each decision will cause corresponding losses. We can use a utility function $l(x)$ to characterize this loss, where $l(x) \neq 0$.

Suppose it is difficult to judge which is more acceptable (characteristic 3), which means we cannot find a valuation function about public acceptance. In that case, scenario S is the "GTPM (general trolley problem model)".

In other words, if the trolley problem is solvable, then we need to be able to find the most acceptable decision, given the limited and inevitable selection and inevitable loss.

The two points in this model are (1) the inevitability of loss and (2) the limit-inevitability of selection. According to this model, we can categorize past work. Research on one decision corresponds to the expansion and division of executive decision types in the trolley problem; research on different scenarios S corresponds to the transformation of the situation from Foot's uncontrollable tram to Thomson's footbridge to ADT; and the construction of different valuation functions regarding the acceptance of the public corresponds to seeking solutions to trolley problems from different ethical theories, such as Kant's deontological ethics, Rawls's theory of justice, and Bentham's utilitarianism.

In a deontological approach, autonomous vehicles should follow principles like "Thou shalt not kill" (Kant), meaning the vehicle should not take actions that harm a human, even if this results in greater overall harm. Rawlsian theory, following a contractualist tradition, protects the most vulnerable and respects equality, refusing to sacrifice one person's interests for another's benefit. The utilitarian approach holds that an action is morally correct if it achieves the greatest benefit for the greatest number.

4. Solutions to the Trolley Problem

It is difficult to reach a unified opinion since different scholars have their own views on dealing with the GTPM. That is the reason why any solution is criticized, which is unacceptable in the autonomous driving field, where the industry is bound to advance. Therefore, someone asked a different question: "Why not prove the trolley problem in the context of ADT does not exist if we cannot find an effective solution?" This path seems to bring light to the GTPM, and researchers generally put forward two solutions:

The first solution acknowledges the existence of the trolley problem in the context of ADT but does not acknowledge the existence of a "most acceptable decision", which means this solution denies that people can rank the solutions from different ethical considerations in the trolley problem in the context of ADT.

The second solution denies the possibility of scenario S directly, thus avoiding answering the trolley problem in the context of ADT. Specifically, this solution can be divided into two different schemes: no general solution to the GTPM and no trolley problem in the field of ADT.

4.1. No General Solution to the GTPM

This scheme argues that it is a moral stricture to require driverless cars based on ADT to solve the trolley problem. Why should we expect driverless cars to solve this problem if human beings have no way to solve the problem [18]?

The problem with this view is that it presupposes that machine learning-based ADT cannot provide a "perfect" solution beyond humans. If "human beings cannot effectively solve the trolley problem", then "ADT cannot effectively solve it", which is a controversial logic. Why should we develop it if ADT cannot solve all kinds of driving problems (including the trolley problem) more effectively than human drivers? That is why the second rule of the Ethics Commission: Automated and Connected Driving report, published in 2017, clearly states the following: "The licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words, a positive balance of risks". After all, there are many examples of automation technology performing as well as or even better than human beings. Therefore, such a scheme is more like an evasion.

4.2. No Trolley Problem in the Field of ADT

Sven Nyholm and Jilles Smids proposed a new idea: we should focus on the differences between the moral dilemmas in the traditional context and the trolley problem in ADT. Sorting out the differences between the two proved that it was meaningless to discuss the trolley problem in the context of ADT [19].

The proponents of this scheme argue that self-driving cars' plight is not a trolley problem but merely an "accident". They have isolated some essential differences between the ethics of accident algorithms for self-driving cars and the trolley problem, which revolve around three main aspects: (i) the overall decision situation and its features; (ii) the role of moral and legal responsibility; and (iii) the epistemic situation of the decision-makers. In detail, the scheme can be divided into five points:

1. The moral accident decision is faced by groups of individuals and multiple stakeholders, while the trolley problem is faced only by one individual;

2. The moral accident decision focuses on prospective decisions/contingency planning, while the trolley problem is immediate;

3. The moral accident decision considers unlimited and unrestricted situational features, while the trolley problem is restricted to a small number of considerations;

4. The moral accident decision takes moral and legal responsibility into account, while the trolley problem does not take responsibility into account;

5. The moral accident decision involves a mix of risk estimations under uncertainty, while the trolley problem considers the facts to be both certain and known.

Nyholm mines the difference between classic scenes and ADT scenes to negate the existence of the trolley problem. However, there are still some problems with the specific application process. On the one hand, the trolley problem's definition considers the classic "dilemma" scene. It is rather unconvincing to deny the existence of the trolley problem in ADT based on this definition. On the other hand, even if we decide that there is only an "accident scene" but no "trolley problem" in the ADT scene, we can still use an "accident scene" as a challenge to ADT, so the problem remains unsolved.

5. Solution Based on "Sufficient Time"

Nyholm's plan teaches us two lessons: first, the trolley problem may not exist in ADT; second, we can construct a comparison system to specifically analyze the different characteristics of the trolley problem in the traditional and ADT scenes, which is beneficial to understand the argument.

ADT's core framework includes perceiving, planning, and controlling [20–22]. In the process of ADT operation, the perceiving system recognizes the scene through different sensors; then, the planning system plans the actions of the self-driving car according to the data provided by the perceiving system. Finally, the controlling system executes the command from the planning system.

In this framework, every operation of each system needs to take a certain amount of time. Based on whether the required time is sufficient or not [23], we can obtain eight scenarios using the grid search method (Table 1).

Table 1. Scenario analysis by the framework of ADT.

Scenarios	Perceiving Time	Planning Time	Controlling Time
1	Insufficient	Insufficient	Insufficient
2	Sufficient	Insufficient	Insufficient
3	Sufficient	Sufficient	Insufficient
4	Sufficient	Sufficient	Sufficient
5	Insufficient	Insufficient	Sufficient
6	Insufficient	Sufficient	Insufficient
7	Insufficient	Sufficient	Sufficient
8	Sufficient	Insufficient	Sufficient

Among them, "sufficient perceiving time" refers to sufficient time for the perceiving system to perceive the current scene through various sensors and digitize it. "Sufficient planning time" means that the planning system has enough time to reconstruct the contemporary scene and make decisions based on the data from the perceiving system. "Sufficient controlling time" means that the controlling system has sufficient time to execute the instructions given by the planning system accurately.

There is a sequence between the three systems. Only after the completion of the current perception can a decision be made, and only after the planning system performs can the controlling system perform. Therefore, scenarios 5, 6, 7, and 8 are not reasonable (the case of cyclic data flow is not considered here). For example, in scenario 7, sufficient planning time and controlling time are required, while sufficient perceiving time is not. However, if the task of the perceiving system is not completed, there is no way to transmit the result to the planning system, so it cannot complete its task. Similarly, the controlling system is unable to complete its job. According to the definition of "sufficient time", we cannot say that the planning and controlling times are sufficient. Therefore, scenario 7 is not valid and realistic.

As for the other scenarios, scenario 1 means that the perceiving system, planning system, and controlling system all have no time to run, which means that the accident occurred very suddenly. The sensor did not finish the work of perception, and the systems did not have time to "think" and make decisions. Then, an accident is bound to occur, and the trajectory of the vehicle will not change. Scenario 2 means that the perceiving system has time to work when the accident occurs, but the planning and controlling systems do not. Although the scene has been perceived and digitized, the planning system cannot "think" and make decisions, so it cannot give any instructions to the self-driving car. In the same way as scenario 1, accidents are bound to occur, and the vehicle's trajectory will not change. Scenario 4 indicates that when an accident occurs, the automatic driving system has enough time to complete perceiving, planning, and controlling, which can avoid the occurrence of an accident. Therefore, accidents cannot be avoided when scenes 1 and 2 occur, and there is no possibility for the subject of responsibility to choose, so there is legal responsibility but no moral responsibility. Scenario 4 is an accident-free case, and there is no need to discuss the so-called moral dilemma.

The most notable scenario is scenario 3, which says that even though there is time for perceiving and planning, there is a lack of sufficient time for effectively controlling driverless cars. Self-driving cars either cannot fully realize all of the instructions issued in the planning process or implement the commands with bias. In this case, losses are sure to occur. (If there is a plan with enough controlling time, it must be considered and implemented first in the planning process.)

However, people hope for a controllable situation on the premise of reducing the accident rate, but the deviation in scenario 3 is unpredictable. Therefore, it would be better to acknowledge the shortcomings of ADT while at the same time using this technology when it is effectively recognized and controlled to keep it in a balanced and controllable state. It is all about people's pursuit of controllability, whether we require the division of accident liability at Level 3 (responsibility controllable); anti-leakage measures of data, such as vehicle driving records (data controllable); or avoidance of the unpredictable deviations (decision controllable) mentioned above. Therefore, for the sake of controllability, the researchers studying the trolley problem believe that it is better to make decisions when considering the existence of losses than to plan a decision scheme that theoretically has no loss but cannot be realized by the final control system, which leads to unknown results.

This default accident scenario with loss satisfies the limit-inevitability of selection, the inevitability of loss, and the complexity of evaluation, which means it is a typical GTPM. In other words, the necessary condition for the emergence of the GTPM in ADT must be to design goals with losses. Let us assume that the controlling time required for the default accident scenario with loss is less than that required for the idealized and lossless scenario. Then, the former approach must be easier to achieve in reality. However, such goals designed without losses will inevitably lead to the recurrence of the trolley problem in ADT. Therefore, one has the following question: Can the autonomous vehicle be designed on the premise of a default accident or risk rate?

6. Limit Feasible Region in Autonomous Driving Technology

Scenarios 1 and 2 lack the time for perceiving or planning, which means we cannot choose the corresponding decision, so there is no trolley problem (limit-inevitability of selection). In scenario 4, we have plenty of time to find a corresponding decision d_i in which $l(x) = 0$, so there is no trolley problem (inevitability of loss). In scenario 3, we face the judgment criterion "most acceptable decision" on the premise that the default loss $l(x)$ exists, which means $l(x) \neq 0$. It is not easy to judge which decision is more acceptable, and one must select the corresponding decision. Such a result satisfies the three characteristics of the GTPM.

However, once the agent of the trolley problem is transferred to the automatic driving system, everything will be different.

The problem is the default condition $\exists\, l \neq 0$. Can the automatic driving system plan with the default loss? It should aim for a zero-accident rate and not presuppose human sacrifice or accidents, although we know that ADT cannot reduce the accident rate to zero. In short, we can accept a self-driving car with a zero-accident rate as its target but an actual accident rate of 0.02%, but it is difficult to accept a self-driving vehicle with a default accident rate of 0.01%. It is not a question of whether the technical design can be achieved but whether the idea can be acceptable.

Google suffered criticism and dissatisfaction from the public when its self-driving cars tried to plan with loss targets, which led to the autonomous driving system choosing to sacrifice "smaller objects" in an accident in 2014. The autonomous driving system prioritized protecting pedestrians and sacrificing "smaller items" in a damaging accident. Similarly, in 2016, Mercedes said it should give priority to the safety of passengers in its cars under the preconditions of loss, which met a strong protest from the public (people arguing that the company had no right to make life choices) until Mercedes re-announced the goal to "avoid 100 percent of accidents" instead of making ethical choices.

The only guidance on the ethics of self-driving cars is from the German Ethics Commission on Automated and Connected Driving (GECACD) office, released by the German Ministry of Transport in June 2017. The GECACD proposes that the introduction of loss presupposition into the automatic driving system is not allowed.

The third rule in the GECACD states the following: "The guiding principle is the avoidance of accidents, although technologically unavoidable residual risks do not militate against the introduction of automated driving if the balance of risks is fundamentally positive". At the same time, rule No. 5 states that the main objective of ADT is to avoid accidents as much as possible when we consider the trolley problem in the context of ADT: "Automated and connected technology should prevent accidents wherever this is practically possible. Based on the state of the art, the technology must be designed in such a way that critical situations do not arise in the first place. These include dilemma situations, in other words, a situation in which an automated vehicle has to "decide" which of two evils, between which there can be no trade-off, it necessarily has to perform".

On the other hand, the guidance also challenges so-called "moral algorithms" in rule No. 8: "They (genuine dilemmatic decisions) can thus not be standardized, nor can they be programmed such that they are ethically unquestionable. Technological systems must be designed to avoid accidents. However, they cannot be standardized to a complex or intuitive assessment of the impacts of an accident in such a way that they can replace or anticipate the decision of a responsible driver with the moral capacity to make correct judgments".

All these examples and guidelines show that it is difficult for the public to accept and for the government to formulate the handling mechanism for the preset loss of the automatic driving system, and they have tried to exclude the preset loss from the design of the automatic driving system.

Going back to the GTPM, the reason why there is no dilemma in the context of ADT is that it is difficult for people to accept the corresponding losses $l(x) \neq 0$ caused by a limited number of decisions d_i. Because of such concerns, people require the automatic driving system to set a zero-risk preset, which is a narrow feasible decision region. However, such a requirement of ADT from the public contradicts the inevitability of losses from the GTPM. If we take the zero-risk or -accident preset in ADT as a GTPM, then it shall correspond to a possible option so that people on both sides will survive in the GTPM. Such a perfect approach shows that ADT's zero-risk or -accident preset is incompatible with the traditional GTPM.

Once we change the context from the traditional scene to the ADT scene, the task of the self-driving car changes from "according to the data provided by the perceiving system, planning of self-driving car's driving path" to "according to the data provided by the perceiving system, planning of self-driving car's driving safe path". Under the premise of insufficient controlling system time, the task of the planning system can never be completed, which rather rules out scenario 2 analyzed by the framework of ADT.

Further, a zero-risk or zero-accident preset in ADT requires the planning system to consider the controlling system's implementation during the decision-making process. Consequently, if the planning system is completed, then the controlling system must be completed as well. These conclusions suggest that scenario 3 is not consistent with the zero-risk or zero-accident preset.

Therefore, we can find that once we accept the zero-risk presupposition of the ADT, there are only two scenarios (Table 2).

Table 2. Scenario analysis by the framework of ADT in zero-risk presupposition.

Scenarios	Perceiving Time	Planning Time	Controlling Time
1*	Insufficient	Insufficient	Insufficient
4*	Sufficient	Sufficient [1]	Sufficient

For scenario 1*, once the time for the controlling system is insufficient, the time for the other two systems is also insufficient (sequentiality). For scenario 2*, once the time for the controlling system is sufficient, the preset of zero loss is satisfied. In this sense, the time for perception, decision-making, and control is sufficient. Therefore, neither scenario 1* nor scenario 4* meets the GTPM according to the definition. In fact, scenario 1* is a totally "uncontrollable" case for ADT, while scenario 4* is a possible zero-loss-option case. If the decision-maker in the GTPM is incapable of doing anything or has a perfect and dominant option, then the GTPM is no longer a dilemma.

7. Free Will and Moral Responsibility in GTPM

Based on the above explanation of the GTPM in the context of ADT, we can further analyze the similarities and differences between the classic and ADT scenarios.

Neither system can find a zero-loss scheme that can be executed (in the classic scenario, the decision system cannot give a zero-loss scheme at any time). The difference is that in the classic trolley problem, the subjects presented with choices are human beings, and they make a choice with their free will or a revisionist free will [24,25]. Free will is often connected with moral responsibility [26–29] and can even be the premise of moral responsibility [30,31]. It is difficult for us to judge which decision is more acceptable because of the moral responsibility (any decision can be accepted), and the complexity of evaluating which decision is more acceptable is the premise of the existence of a "dilemma".

ADT aims to avoid accidents all the time. In scenario 3, the system is limited by this goal and cannot find a decision d_i that makes $l(x) = 0$. Therefore, it can only be in the state of "no decision". This "no decision" is not a "no choice" based on free will but a forced choice in a state of non-freedom, that is, a forced choice that was designed to prevent making decisions in situations when $l(x) \neq 0$ in an automatic driving system. Therefore, there is no need to discuss the trolley problem in the context of ADT.

In fact, a similar idea has already been suggested by Johannes Himmelreich, and we offer a more detailed argument [32].

Of course, it is still necessary to investigate the relevant legal liability after an accident. In the state of "no free will", self-driving cars tend to follow the established route when accidents cannot be avoided. The question is: Who caused the self-driving vehicle to fall into this situation of "certain loss"?

If the loss of the self-driving car cannot be avoided due to human error, then the responsible party is the person who caused the mistake. Suppose the perceiving, planning, or controlling system is wrong due to a quality problem; thus, the loss cannot be avoided. In that case, the responsible party shall be the manufacturer or technology provider of the vehicle. If unavoidable losses are caused by natural force majeure, such as earthquakes and floods, then relevant social assistance should be sought. For a man-made force majeure, such as unqualified works leading to inevitable loss, the main body that caused the force majeure shall be investigated for responsibility.

There are many specific cases. We may need an independent public sector agency (for instance, a Federal Bureau for the Investigation of Accidents Involving Automated Transport Systems or a Federal Office for Safety in Automated and Connected Transport) to deal with responsibility arising from self-driving cars (rule No. 8 in the GECACD).

8. Conclusions

The difference between the classic trolley problem and the trolley problem in the context of ADT is that the former allows the driver to make decisions under the premise of loss, while the latter does not. It is because of this limitation that the decision-making system of the latter is trapped in a limited feasible region, and the agent of decision-making, the automatic driving system, is always in a state of no discretion. In this situation, the existing condition of the GTPM is disintegrated, and the dilemma no longer exists.

Therefore, this paper concludes that the trolley problem does not exist in the field of autonomous driving, as autonomous driving systems cannot recognize moral dilemma scenarios and thus cannot make ethical decisions based on such recognition. The illusion that the trolley problem applies to autonomous driving arises mainly from three factors:

Confusion between thought experiments and the real world: The former is a static process where we have ample time to determine the characteristics of the current scenario and make judgments, while the latter is an emergency where autonomous driving systems lack sufficient time for judgment.

Neglect of the prerequisites of moral dilemmas: An important prerequisite of moral dilemma scenarios is that loss is inevitable, and all decisions must be made based on this premise. However, in real-world scenarios, no one would allow autonomous vehicle manufacturers to construct path planning systems based on the premise of “inevitable loss”.

Treating continuous possibilities as discrete: In the classic trolley problem, the options are limited (usually only two), and we can make utility calculations between different possible options. However, real life is continuous, and the path planning of autonomous driving systems has infinite possibilities. Therefore, their goal can only be to “minimize loss”, rather than calculating losses for discrete options and then invoking a “trolley problem algorithm” to solve the issue.

One may argue that the conclusion in this paper is a kind of evasion. One may claim that the statement about the nonexistence of a dilemma does not help solve the problem. However, such an argument misses the target. This paper does not state that there will never be any GTPM-like dilemma in ADT, whether in a theorizing or thought experiment scenario. Actually, scenario 3 in Table 1 is a typical trolley problem case. It is the unacceptance of default risk or accident preset by the public that makes scenario 3 impossible. The logic of this paper’s argument relies on the premise of actual policy in reality rather than pure concept analysis.

One may further argue that the old philosophical discussions on emerging AI technological issues such as ADT are hardly suitable. Such an opinion overlooks the reflection and forward-looking role of philosophy. The reflections of philosophy about emerging science and technology issues are beneficial to discovering potential problems such as dilemmas or paradoxes. Moreover, these discussions do not only stay at the level of the imagination but attempt to analyze the causes and find solutions through conceptual analysis and thought experiments. At least some misconceptions and ambiguities can be clarified through such a process. On the other hand, philosophical discussions on ongoing technological issues are supposed to avoid overcomplicating the problems. For example, when considering ADT-related problems, it is not appropriate to take ADT as some strong AI, which may be regarded as complete moral and epistemic agents in the future. Such considerations are not attempts to solve the problem under the current framework of ADT. Whether strong AI can be fully realized or ADT should be considered as a moral agent is a controversial topic [33–36], and introducing such arguments to current ADT may lead to excessive worries. After all, ADT is only a working machine designed to complete a relatively single task over a long period of time into the future.

Although this paper refutes the presence of the trolley problem in the context of autonomous driving, it acknowledges the ethical issues arising from path planning in autonomous driving systems. These issues, while not ethical dilemmas in the traditional sense, are significant in determining whether individuals choose to adopt autonomous driving technologies. Our future research endeavors will focus on exploring the nature of these issues; examining their distinctions from traditional moral dilemmas; discussing the differences between ideal scenarios, thought experiments, and real-world situations; and investigating how people address these issues in practical contexts.

Author Contributions: Conceptualization, H.Z.; methodology, H.Z.; validation, D.W.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z. and D.W.; supervision, H.Z.; project administration, H.Z.; funding acquisition, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hunan Normal University, RESEARCH ON THE INTERPRETABILITY OF AI, grant number P2021001.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kriebitz, A.; Max, R.; Lütge, C. The German Act on Autonomous Driving: Why ethics still matters. *Philos. Technol.* **2022**, *35*, 29. [CrossRef]
2. Caro-Burnett, J.; Kaneko, S. Is society ready for AI ethical decision making? Lessons from a study on autonomous cars. *J. Behav. Exp. Econ.* **2022**, *98*, 101881. [CrossRef]
3. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The moral machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef]
4. Awad, E.; Dsouza, S.; Shariff, A.; Rahwan, I.; Bonnefon, J.F. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 2332–2337. [CrossRef] [PubMed]
5. Jean-François, B.; Azim, S.; Iyad, R. The social dilemma of autonomous vehicles. *Science* **2016**, *352*, 1573–1576.
6. Azim, S.; Jean-François, B.; Iyad, R. Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* **2017**, *1*, 694–696.
7. Marchant, G.E.; Lindor, R.A. The coming collision between autonomous vehicles and the liability system. *Santa Clara L. Rev.* **2012**, *52*, 1321.
8. Douma, F.; Palodichuk, S.A. Criminal liability issues created by autonomous vehicles. *Santa Clara L. Rev.* **2012**, *52*, 1157.
9. Xiao, S.; Cao, J. On the Civil Liability of Artificial Intelligence. *Sci. Law* **2017**, *35*, 166–173.
10. Geisslinger, M.; Poszler, F.; Lienkamp, M. An ethical trajectory planning algorithm for autonomous vehicles. *Nat. Mach. Intell.* **2023**, *5*, 137–144. [CrossRef]
11. Zhao, L.; Li, W. "Choose for No Choose"—Random-Selecting Option for the Trolley Problem in Autonomous Driving. In *LISS2019: Proceedings of the 9th International Conference on Logistics, Informatics and Service Sciences*; Springer: Singapore, 2020; pp. 665–672.
12. Wu, S.S. Autonomous vehicles, trolley problems, and the law. *Ethics Inf. Technol.* **2020**, *22*, 1–13. [CrossRef]
13. Philippa, F. The problem of abortion and the doctrine of double effect. *Oxf. Rev.* **1967**, *5*, 5–15.
14. Judith, J.T. Killing, letting die, and the trolley problem. *Monist* **1976**, *59*, 204–217.
15. Derek, L. A Rawlsian algorithm for autonomous vehicles. *Ethics Inf. Technol.* **2017**, *19*, 107–115.
16. Gray, M. Moral machines. *New Yorker*. 2012, p. 24. Available online: https://www.newyorker.com/news/news-desk/moral-machines (accessed on 21 August 2024).
17. Wendell, W.; Colin, A. *Moral Machines: Teaching Robots Right from Wrong*, 1st ed.; Oxford University Press: New York, NY, USA, 2008.
18. Jianwu, L. Capacity Difference and Responsibility Difference: On Possibility of Driverless Vehicle as Ethical Subject. *Soc. Sci. Yunnan* **2018**, *4*, 15–20+186.
19. Sven, N.; Smids, J. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory Moral Pract.* **2016**, *19*, 1275–1289.
20. Keqiang, L. Key topics and measures for perception, decision-making and control of intelligent electric vehicles. *Sci. Technol. Rev.* **2017**, *14*, 85–88.
21. Zhang, X.; Gao, H.; Zhao, J.; Zhou, M. Overview of deep learning intelligent driving methods. *J. Tsinghua Univ. (Sci. Technol.)* **2018**, *58*, 438–444.
22. Basye, K.; Dean, T.; Kirman, J.; Lejter, M. A decision-theoretic approach to planning, perception, and control. *IEEE Expert* **1992**, *7*, 58–65. [CrossRef]
23. James, B.; Yoshua, B. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
24. Stephen, K. Aborting the Zygote Argument. *Philos. Stud.* **2012**, *160*, 379–389.
25. Manuel, V. *Building Better Beings: A Theory of Moral Responsibility*, 1st ed.; Oxford University Press: New York, NY, USA, 2013.
26. Fischer, J.M. Recent work on moral responsibility. *Ethics* **1999**, *110*, 93–139. [CrossRef]
27. Richard, D. *The Non-Reality of Free Will*, 1st ed.; Oxford University Press: New York, NY, USA, 1990.
28. Laura, E. *Free Will: A Philosophical Study*, 1st ed.; Routledge: New York, NY, USA, 2018.
29. Saul, S. *Free Will and Illusion*, 1st ed.; Oxford University Press: Oxford, UK, 2000.
30. Susan, W. *Freedom within Reason*, 1st ed.; Oxford University Press: New York, NY, USA, 1990.
31. Alfred, M. *Free Will and Luck*, 1st ed.; Oxford University Press: Oxford, UK, 2006.
32. Johannes, H. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory Moral Pract.* **2018**, *21*, 669–684.
33. Colin, A.; Gray, V.; Jason, Z. Prolegomena to any future artificial moral agent. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 251–261.
34. Frances, G.; Keith, M.; Marty, W. The ethics of designing artificial agents. *Ethics Inf. Technol.* **2008**, *10*, 115–121.
35. Patrick, C.H. Artificial moral agents are infeasible with foreseeable technologies. *Ethics Inf. Technol.* **2014**, *16*, 197–206.
36. Awad, E.; Dsouza, S.; Bonnefon, J.F.; Shariff, A.; Rahwan, I. Crowdsourcing Moral Mach. *Commun. ACM* **2020**, *63*, 48–55. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.