Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill

Rolf Johansson, Jonas Nilsson

▶ **To cite this version:**

Rolf Johansson, Jonas Nilsson. Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill. Workshop CARS 2016 - Critical Automotive applications : Robustness & Safety, Sep 2016, Göteborg, Sweden. hal-01375606

HAL Id: hal-01375606
https://hal.science/hal-01375606

Submitted on 3 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill.*

Rolf Johansson, *Member, IEEE,* and Jonas Nilsson

***Abstract*— Many claim that the "Trolley problem" is hindering the introduction of self-driving cars. Self-driving cars must, as all safety-critical products, be designed such that the probability of morally hard ("trolley") situations is acceptably low. In this paper we argue that the introduction of self-driving cars can solve this ethical dilemma. In short the solution to the trolley problem is that a self-driving car must be able to estimate its own operational capability for handling surprising situations, and adjust its own tactical behavior accordingly. By limiting the risk for the case of not being able to handle all surprising events in a similar way as for other safety goals today, the remaining risk for the trolley problem can be argued as low as any other acceptable risk of vehicle E/E implemented functionality.**

I. Introduction

The rapid development of autonomous vehicles has brought increased attention to ethical considerations of this new technology. One common headline present in various media is that "Self-driving cars will need to be programmed to kill their owners, academics warn, and people will have to choose who will die".

How should we approach and solve the trolley problem for autonomous vehicles? It is straightforward to construct an example when an autonomous vehicle is faced with a similar choice, e.g. steering left will cause one accident while steering right will cause another. Deployment of autonomous vehicles will without doubt require addressing this problem.

In this paper we present an approach for solving the autonomous vehicle version of the trolley problem. This is done using the following outline. Section II gives some historical background to ethical dilemmas. In Section III we define the term "safe" while Section IV describes the decision-making hierarchy in an autonomous vehicle. The main contribution is presented in Section V, where we describe how to argue that an autonomous vehicle is safe and thus disarming the trolley situation. Finally, concluding remarks are stated in Section VI.

II. Ethical Dilemmas

The question of what is right and wrong has always created a debate in society, especially when it comes to how to act in a scenario when there is no obvious preferred choice. An often cited ethical dilemma, originally formulated by Philippa Foot, is called 'the trolley problem', [1]. It has then become the standard reference following the analysis in the mid-seventies, [2].

The given assumption is that something surprising happens and then there are two choices on how to act. If no active choice is made then some persons are killed, and if an active choice is made the consequence is that other but fewer persons die. The dilemma is named from the scenario of a trolley running down a track unable to brake, approaching a fork point. You are beside the track having the time to reach a lever which can enable you to make the trolley change track. If you do not act, five people will be killed. If you pull the lever and make the trolley turn, another but single person will die.

The problem is to illustrate the conflict between what is called utilitarianism and deontology, respectively. In the former case you try minimize the total harm (here minimize the number of deaths), and in the latter case you avoid to do things that always are wrong (here you avoid to actively kill a certain person).

The problem shows up in several variants, where for example the active choice leads to the killing of your child or of yourself, still this is done to save a larger number of lives. In the context of self-driving vehicles, instead of a running trolley, there is for example a car being surprised by a situation where it either continues in its lane killing a school class running out in front of the car, or takes evasive action off the road killing the driver. The choice is in the hands of the autopilot.

As the plans for self-driving cars has become more and more concrete, and test driving of these is a reality, the question has been more frequently discussed. The perspective is mainly of philosophy, psychology and insurance economics, and then referenced by journalists and other actors on the internet.

In 2013 there was a joint paper from different disciplines and different countries asking if autonomous vehicles would be ready for utilitarianism, Bonnefon et al., [3]. They discuss to what extent people are willing to accept a consequent utilitarian behaviour of a self-driving car, i.e. always minimizing the number of deaths even if this implies killing the person inside the self-driving car. When Google's self-driving car became more visible, the above article was discussed in the newspaper The Independent, [4], addressing a broader public with this question.

Almost at the same time as the article by Bonnefon et al, in October 2013, the magazine The Atlantic had an article about the need to discuss ethics of autonomous cars, [5]. Also here the trolley problem is discussed. In addition to that, the possible dilemmas that may occur if a self-driving car is

* Research has been supported by the Swedish government agency for innovation systems (VINNOVA) in the FUSE project (ref 2013-02650).

Rolf Johansson is with the SP, Technical Research Institute of Sweden, Boras, Sweden; e-mail: rolf.johansson@sp.se.

Jonas Nilsson is with the Volvo Car Corporation, Göteborg, Sweden; e-mail: jonas.nilsson@volvocars.com.

driving 100% according to the traffic rules were highlighted as well as the problem of human drivers provoking the self-driving vehicles. The conclusion of the authors is that all these ethical problems need to be discussed more extensively.

During 2015, the question has been addressed in many places, here just referencing a few, e.g. [6], [7], [8]. In the latter paper the utilitarian approach is taken one step further, introducing cooperative driving, letting the cars collectively analyse and decide how to minimize the number of persons to kill.

All the above publications have in common that they claim the trolley problem is important to address when designing autonomous vehicles. It is easy to get the impression that they regard the trolley problem not only as a constructed problem used to discuss ethical dilemmas in general, but also as a real and important problem to address in the context of self-driving vehicles. Still, it is hard to say that the trolley problem is seen as a real dilemma for the majority of manual drivers. In the following sections we discuss how the trolley problem in the context of manual and autonomous driving, respectively, relate to each other, and the implications on how we need to program the self-driving vehicles to make them safe.

III. What is Safe

When arguing road traffic safety today, the Vienna Convention, [9], is an often cited international agreement telling that the driver "shall at all times be able to control his vehicle". This convention is not ratified by all nations, and among its signees the implementation in national law and regulations are made differently. Still we can use this as the basis in an argumentation for how to achieve safe road traffic: We ask each driver to control her vehicle in such a way that all regulations are met, including driving safely. If all drivers fulfill such a requirement, there will be no accidents (following by definition from 'driving safely').

If an accident occurs, we search for the responsible. In many cases a driver is found guilty of reckless driving. If for example a car is hitting a child suddenly running out in a street in a residential area, the driver cannot blame the child. As a driver you are responsible to have enough imagination to foresee this possibility, and adjust your driving accordingly. This is why the trolley problem is not very relevant in reality for a majority of drivers. To our knowledge, the trolley problem is not addressed at all in any driver instruction textbooks. On the other hand, there is very much focus in driver education on how to avoid accidents by constantly planning for surprising events. From the societal point of view, we put a lot of responsibility on the drivers. In order to receive a driver's license, you need to show that you can control the vehicle, even in the case of surprising events.

In some cases no driver is found responsible for a severe road accident. It is hard to get an exact number, but if we add human lapses, errors and violations, together they are believed to cause significantly more than 90 percent of severe accidents, [10], [11]. If the reason for an accident is that the vehicle is not behaving according to what the driver should expect, the vehicle manufacturer (OEM) can be found liable (product liability).

For many years, the amount of advanced driver assistance systems (ADAS) in road vehicles, have increased. These are designed to reduce the amount of accidents caused by human errors and lapses, by for example automated emergency braking or corrective steering. Long before the ADAS term was invented, there were electronic/electrical (E/E) systems in many cars having an impact on the driver's capability to control the vehicle. An early example is the anti-lock brake system (ABS). The promised functionality is to adjust the braking request from the driver's pedal down to a level where the wheels will not lock. For the driver to control her ABS-equipped vehicle, the instruction is to press the brake pedal hard enough (too hard is not a problem any longer). However, as we give the opportunity for the E/E system to reduce the brake force, even if this is on a sub-second scale, we have to make sure that this is not made in a faulty way where the braking force is reduced too much for a too long time, eventually causing an accident.

In order to restrict dangerous faulty behavior of E/E controlled functionality, the automotive community has agreed to follow a standard on functional safety, ISO 262626, [12]. Most industries (if not all) developing safety-critical products have a similar approach, e.g. nuclear, avionics, railway. Self-driving cars belong to the automotive domain and consequently the terminology of ISO26262 is used for the remainder of this paper. Nevertheless, we remark that the underlying argumentation is independent of domain.

In this standard it is agreed how to assess potential risks, and what countermeasures that are regarded as sufficient to reduce the risks to an acceptable level. As for any functional safety standard, risk is here a measure taking both severity and frequency into account. This means that if either the severity is low enough or the hazardous event is found to be too improbable, there is no need for dedicated risk-reducing measures. All hazardous events that need risk-reducing measures are assigned an automotive safety integrity level (ASIL).

In the ISO 26262 terminology, extremely unlikely events are said to be of frequency E0 (incredible). An example given in this standard is the situation where: "a vehicle involved in an incident which includes an airplane landing on a highway". This means that the designer of certain E/E functionality should make sure that there is no faulty behavior causing fatal injuries even if the frequency of the critical situation is *very low* (E1), but they do not need to consider faulty behavior that is only causing fatal injuries in *incredible* situations (E0), i.e. E0 situations has no ASIL.

Safety goals are assigned to all hazardous events with an ASIL. A safety goal is a vehicle-level requirement stating for instance that "the vehicle shall stay on the road".

Four things are now needed for the vehicle to be considered *safe*:

1. All hazardous events are identified
2. All hazardous events have a correct ASIL
3. The safety goals cover all hazardous events with an ASIL
4. The safety goals are fulfilled.

By following this standard, an OEM can argue internationally about product liability for severe accidents. As long as every hazardous event including all situations with very low probability has been considered, we can claim the E/E functionality to behave safely if necessary risk reduction measures have been applied. The limit between *very low probability* and *incredible* defines the limit for product liability regarding road safety, according to ISO 26262.

When making a vehicle more and more autonomous, also more and more of the responsibility for driving safely is transferred from the driver to the vehicle. If the driver is told in the owner's manual, that a certain vehicle is capable of driving fully autonomously on certain specified roads, this implies that in order to be functional safe, all possible risks have to be assessed and addressed. As all the driving functionality is transferred from the driver to the vehicle, the responsibility for driving safely is also transferred to the vehicle.

We can conclude that proving that an autonomous vehicle is behaving safely in traffic, can be argued by means of functional safety. If we can show that an autonomous vehicle is functional safe, we know that it will behave safely on the roads. This follows by definition from the functional scope of the autonomous vehicle; to at all times be able to control the vehicle safely. Any deviation from this would imply that the vehicle is not functional safe.

IV. The decision Hierarchy

A major difference between an autonomous vehicle and an ADAS functionality is the responsibility for the former to plan the driving, solving problems proactively and not only reactively, as may be the only alternative for the latter. In a manually driven vehicle, the driver is responsible for all decisions on strategical, tactical and operational levels, respectively. When introducing ADAS to assist the driver, this is mainly to compensate driver errors and lapses on the operational level. What vehicle speed to aim for, or what distance to try to keep to vehicles and pedestrians, are examples of tactical choices. We expect the manual driver to perform such decisions, and then operational level E/E implemented functionality can assist to make the vehicle conform to the tactical decisions. Another example of a tactical decision is whether to start an overtaking maneuver during certain conditions. In the community of autonomous vehicles it has for a long time been evident that the step towards autonomy implies taking the responsibility for the tactical decisions, see e.g. [13], [14], [15].

In a sense, the operational task for an autonomous vehicle may become easier to fulfill than for a manually driven vehicle. This is because in the case of a manually driven vehicle, the ADAS functionality should be able to handle all situations regardless of the tactical decisions of the driver; let them be good or bad. In the case of autonomous vehicles, the operational algorithms only have to deal with the situations possible to appear given the tactical decisions of that vehicle. This means that if the tactical choices are 'clever', then the operational task is easier to perform, at least from a safety perspective.

So what is then a 'clever' tactical choice? One way to answer this is by saying that this is something that is possible to handle by the operational capabilities. Another way is to observe that this is to a certain extent what we address in the regulations today when saying that the manual driver should not perform reckless driving. Starting an overtaking maneuver without confidence that it can be performed safely, is an example of a tactical choice that is considered not so clever. The same goes for not adjusting the target speed in a situation where a playing child can surprisingly run out in the street of a residential area, or to limited sight ahead due to bad weather conditions.

We can say that it is possible to balance the responsibility between the tactical choices and the operational capabilities. The more offensive (optimistic) tactical profile, the harder is the task on the operational level to guarantee safety. One way to solve this equation is to adjust the tactical profile to the current operational capabilities. How fast the operational capability is to react on surprises, sets the limit for how severe surprises the tactical choices may lead to. This means that for an autonomous vehicle a 'clever' and safe tactical decision pattern will not be to mimic an experienced driver, but to balance the current operational capabilities of itself. In some situations this will likely imply more conservative driving, but in other situations a less conservative style, compared to human drivers. The important thing is that the decision hierarchy is consistently implemented in such a way that the tactical choices always guarantee the operational tasks to be solved by the vehicle. How to argue for such completeness and consistency, and how this solves the trolley problem, are further elaborated in the next section.

V. Safety Argumentation

Let us assume that we have identified all hazardous events with an ASIL and assume for the sake of simplicity that we have formulated two generic safety goals for an autonomous vehicle: 'stay on the road' and 'avoid collisions', respectively. In reality there will be more precise safety goals stating for instance what type of collision must be avoided with what ASIL. For the importance of considering different tolerance margins of every safety goal, see for example [16]. This means that the two above generic safety goals will be refined to four safety goals (one for each ASIL value) for each object category (vulnerable road users, personal cars, trucks, elks, stationary large object, etc, or any favorite classification of whom to avoid collisions with).

According to the discussion in Section III, fulfilling these would make us master the trolley problem. If we can guarantee never to collide and never to run off the road, this means that the catastrophic scenarios of the trolley problem will be avoided.

The above safety goals should be fulfilled at all times by the autonomous vehicle. To show that this is the case, we refine and allocate safety requirements on a more and more detailed architecture. Following the pattern of a functional safety concept architecture of [15], we get a generic pattern of division of responsibilities between an environment perception (EP) block and a decision and control (DC) block as shown in [17]. In order to disarm the trolley problem, we then need to further refine the safety requirements inside the

DC block showing the responsibilities of the tactical and the operational blocks, respectively.

The trick is now to consider all the different ASIL versions of the operational functional safety requirements (FSR) and formulate them with such tolerance margins that they all can be shown to be fulfilled. In other words, we adjust the request to the operational block to the capability we can guarantee. Then we adjust the tactical FSR such that they do not require any higher operational performance than we just found possible to assess. If we then can implement and asses such FSR on the tactical block, we can claim that we have disarmed the trolley problem for the autonomous vehicle. This problem is always solvable. If the operational capability is reasonably high, the tactical choices can imply a rather 'normal' driving style. If the operational capability is lower, the driving style will more conservative, but never unsafe.

What we have done is to adjust the tactical decisions such that if we will get surprised in the next instant, we have time and capacity on the operational level to handle every surprise safely. This is precisely what an educated manual driver aim for. The difference is that a human driver makes mistakes both in judging her own operational capability, and in keeping alert while driving. The advantage of an autonomous vehicle is that it can stay alert without getting tired or distracted, and if we can make it capable to judge the differences in operational capabilities in run-time, we can program it to always make 'clever' tactical decisions.

A conclusion of the above reasoning is that while the trolley problem can be fully solved for an autonomous vehicle, it is a real dilemma for ADAS systems in manually driven cars. Today OEM vehicle manufacturers tend to take the position of deontology, i.e. not making anything at all actively is always to prefer in case of a conflict. The underlying argumentation is that ADAS is only assisting the manual driver, and the driver is fully responsible. But in contrast to Bonnefon et al., we claim that it is only for manually driven cars, and not for self-driving cars, this question needs to be discussed. The motivation for this difference is that autonomous vehicle can prove itself to plan its driving in such a way that the remaining risk of unsafe surprising situations is acceptable low, while this is far from true for manually driven vehicles where the ADAS systems may be surprised by impossible situations as often as the manual drivers are making severe mistakes.

VI. Conclusion

Some claim that the trolley problem is a serious concern when deploying autonomous vehicles. This particular ethical dilemma arises when someone (e.g. a self-driving vehicle) as a consequence of a surprising situation is faced with the choice between two catastrophic events.

Self-driving cars must, as all safety-critical products, be designed such that the probability of morally hard ("trolley") situations is acceptably low. There is an internationally agreed risk level for accidents caused by the intelligent vehicle functionality, documented in the ISO 26262 standard. This can be applied for determining the risk assessment needed to master the trolley problem, as all other road traffic risks.

In this paper, we present a functional safety approach to disarm the trolley problem completely for self-driving vehicles. We argue that the self-driving vehicle shall have the responsibility not to get surprised in a way that the trolley problem can show up.

A key enabler to disarm the trolley problem is the ability of the self-driving vehicle to estimate its own operational capability for handling surprising situations, and adjust its own tactical behavior accordingly. Tactical decisions regarding things like vehicle speed, distance to surrounding objects, when to overtake etc, are adjusted according to its current operational capabilities, e.g. confidence from sensing systems and the brake capacity. This capability to adjust tactical decisions to operational capability guarantees a safe disarmament of the trolley problem.

References

[1] Philippa Foot, "The problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices", Oxford Review, Number 5, 1967.

[2] J. J. Thomson, "Killing, Letting die, and the Trolley Problem", The Monist 204-17, 1976.

[3] J-F Bonnefon, A. Shariff, I. Rahwan, "Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?", arXiv:1510.03346v1 [cs.CY] 12 Oct 2013.

[4] A. Griffin, "Self-driving cars will need to be programmed to kill their owners academics warn and people will have to choose who will die", The Independent, October 27, 2015.

[5] P. Lin, "The Ethics of Autonomous Cars", The Atlantic, October 8, 2013.

[6] M. Windsor, "Will your self-driving car be programmed to kill you if it means saving more strangers?", June 15, 2015

[7] MIT, "Why Self-Driving Cars Must Be Programmed to Kill", MIT Technology Review, October, 2015

[8] D. Weinberger, "Should your self-driving car kill you to save a school bus of full of kids", Digital Trends, October 27, 2015.

[9] Economic Commission for Europe, Inland Transport Committee, "Convention on road Traffic", done at Vienna on 8 November 1968.

[10] US Department of Transportation, National Highway Traffic Safety Administration, "National Motor Vehicle Crash Causation Survey", Report to Congress, July 2008.

[11] S. Forward, "Driving Violations", Doctoral Thesis, Uppsala University, 2008.

[12] ISO, "International Standard 26262 Road vehicles -- Functional safety", November 2011.

[13] R. Sukthankar, "Situation Awareness for Tactical Driving", Ph.D. thesis, Robotics Institute, Carnegie Mellon University, USA, January 1997.

[14] T. X. P. Diem and M. Pasquier, "From Operational to Tactical Driving: A Hybrid Learning Approach for Autonomous Vehicles", 2008 10th Intl. Conf. on control, Automation, Robotics and Vision, Hanoi, Vietnam, December 2008.

[15] S. Behere and M. Törngren, "A Functional Architecture for Autonomous Driving", Proceedings of the First International Workshop on Automotive Software Architecture - WASA '15, Montréal, Canada, May 2015.

[16] R. Johansson, "The Importance of Active Choices in Hazard Analysis and Risk Assessment", Critical Automotive applications: Robustness & Safety (CARS), Paris, France, 2015.

[17] R. Johansson, J. Nilsson, "The Need for an Environment Perception Block to Address all ASIL Levels Simultaneously", Workshop on holistic interfaces for environmental fusion models, Intelligent Vehicles Symposium, Gothenburg, Sweden, 2016.