Trolled by the Trolley Problem

On What Matters for Ethical Decision Making in Automated Vehicles

Alexander G. Mirnig
University of Salzburg
Salzburg, Austria
alexander.mirnig@sbg.ac.at

Alexander Meschtscherjakov
University of Salzburg
Salzburg, Austria
alexander.meschtscherjakov@sbg.ac.at

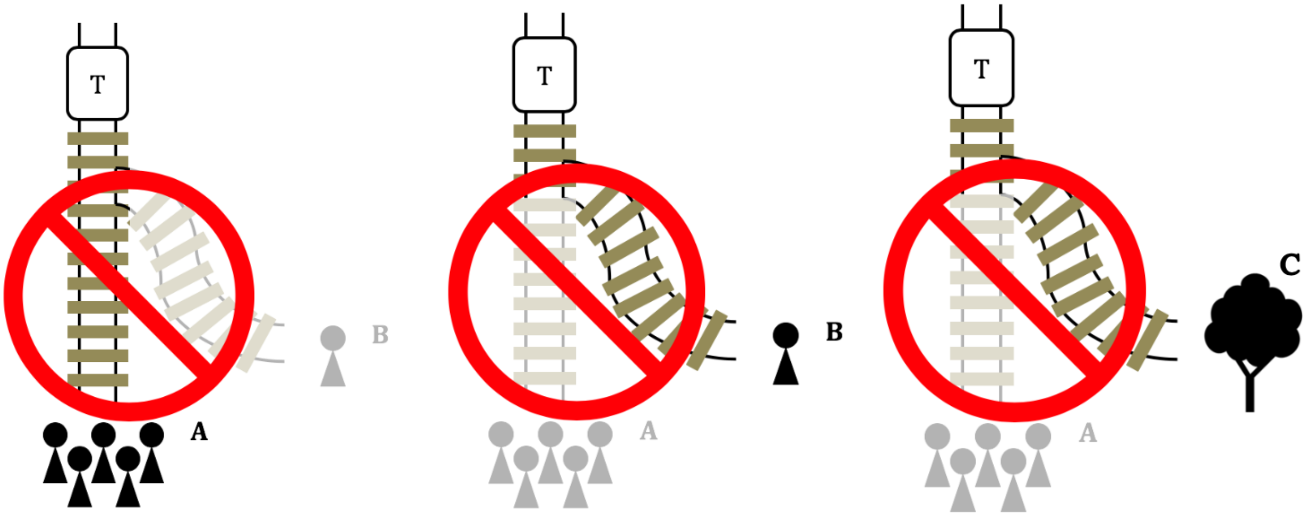

Figure 1: The trolley problem is unsolvable by design. It always leads to fatal consequences.

ABSTRACT

Automated vehicles have to make decisions, such as driving maneuvers or rerouting, based on environment data and decision algorithms. There is a question whether ethical aspects should be considered in these algorithms. When all available decisions within a situation have fatal consequences, this leads to a dilemma. Contemporary discourse surrounding this issue is dominated by the trolley problem, a specific version of such a dilemma. Based on an outline of its origins, we discuss the trolley problem and its viability to help solve the questions regarding ethical decision making in automated vehicles. We show that the trolley problem serves several important functions but is an ill-suited benchmark for the success or failure of an automated algorithm. We argue that research and design should focus on avoiding trolley-like problems at all rather than trying to solve an unsolvable dilemma and discuss alternative approaches on how to feasibly address ethical issues in automated agents.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models.*

KEYWORDS

Automated Vehicles; Trolley Problem; Dilemma; Ethics.

ACM Reference Format:
Alexander G. Mirnig and Alexander Meschtscherjakov. 2019. Trolled by the Trolley Problem: On What Matters for Ethical Decision Making in Automated Vehicles. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3290605.3300739

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300739

1 INTRODUCTION

As automotive technology grows in maturity, so do our expectations in said technology. Vehicles nowadays are not only improving on a mechanical level (being lighter, faster, or safer in case of an accident) but possess a variety of assistive functions as well, with some of them assisting with or even entirely performing individual driving tasks.

Challenges in (partially) automated driving[1] for HCI research and interaction design have been discussed lately (e.g., [8, 10, 28]). While there are also expectations related to added comfort and more efficient use of transit times [3, 19], the primary expectation in automated vehicle technology is related to safety [30]. By eliminating human error from the on-road equation, it is expected to eventually get rid of a – if not *the* – primary cause of on-road accidents today. After all, a machine will never get tired, be negatively influenced by medication or alcohol, or get distracted. Given sufficient computational power and input about the driving environment (be it via sensors or vehicle-infrastructure communication), the vehicle can also be expected to nearly always make the "right" decision in any known situation, as it might not be subject to misjudgement or incomplete information due to limited perception.

This raises an old question within a new context: What shall a deterministic computer do, when confronted with morally conflicting options? This is also know as an *ethical dilemma* often presented as the so-called "trolley problem". In a nutshell, a trolley problem refers to a situation in which a vehicle can take a limited number of possible actions, all of which result in the loss of human life. This means that whichever action is taken, the vehicle has "decided" to take a human life and in creating the vehicle, we have allowed a machine to take this authority over human life, which seems intuitively hard to accept.

In this paper, we argue that the trolley problem serves a very important social function, while also being a poor benchmark for the quality and/or performance of a machine's decision making process. We do this by providing background information on the family of dilemmas the trolley problem belongs to and the purpose of such dilemmas within philosophical discourses. We then outline what a solution to such a dilemma would entail and why these are not suitable for decision making within the automated vehicle context. We introduce recommendations on how to treat the problem in relation to automated vehicles and argue that the trolley problem is unsolvable by design and that, thus, emphasis should be put on avoiding a trolley-dilemma in the first place. Finally, we present a discussion of suitable benchmarks for automated vehicles that appropriately consider the legal and moral perspectives in a vehicle's decision making process in relation to the specific context and outline a design space that can be addressed by the HCI community.

In tradition with, for example, Brown et al. [7] who raised provocative questions for ethical HCI research, we want to question whether the omnipresent trolley dilemma is the best way to deal with the subject of moral decisions of automated vehicles. We intend to help HCI community members (including individuals or bodies involved in regulation, developers and interaction designers) to build a nuanced understanding of the complexities of the trolley problem.

2 RELATED WORK

With automated vehicle technology continuing to be on the rise, discussions about the trolley problem have also gained increasing attention in public media and discourse [2, 37]. In the scientific community, the modern incarnation of the trolley problem are often attributed to Philippa Foot and Judith Jarvis Thomson. The former presented the trolley problem as one example variant in her 1967 paper about abortion dilemmas [12], whereas the latter published a rigorous analysis specifically about the trolley problem [38].

Within automotive-related domains, such as robotics, software engineering, and HCI, the trolley problem has received significant attention, particularly in relation to automated vehicles (see e.g., [33]). For example, in human-robot interaction first steps towards developing the field of Moral HRI have been suggested to inform the design of social robots. Malle et al. [25] researched how people judge moral decisions of robots in comparison to human agents. They found that participants expected robots to follow a utilitarian choice more often than human agents. Above that, their research suggest that people blame robots more often than humans, if the utilitarian choice was not taken.

Goodall [15] discussed ethical decision making of machines for the automated vehicle context providing a basis for ethical discussions. The MIT Media Lab set up the "Moral Machine[2]", an on-line interactive database with different moral traffic dilemmas, where the user can provide input on how the vehicle should respond in these situations and how they judge these situations from a moral perspective. Bonnefon et al. [5] pointed out a social dilemma when using the trolley problem as means to inform decision making for autonomous vehicles. They found that participants would

[1] The SAE J3016 standard [34] distinguishes between six levels of driving automation from level 0 (i.e. no automation) to 5 (i.e. full automation). In this paper we address automated vehicles of level 3 (i.e. conditional automation) or higher, since then the vehicle is required to make its own decisions without any driver intervention acting as an autonomous agent. Henceforth, we use the term 'autonomous' to highlight the capability of an agent (may it be vehicle or human) to act independently. For a vehicle, that also includes the capability to transfer control to the driver as it is the case in SAE level 3.

[2] http://moralmachine.mit.edu

approve utilitarian algorithms in automated vehicles in general, but when it comes to sacrificing their own live they would demand automated vehicles to protect the passengers lives at all costs. Thus, participants were against enforcing utilitarian regulations predicting not to buy such vehicles. In conclusion, it is argued that a regulation for utilitarian algorithms may postpone the adoption of such vehicles, which in turn paradoxically decreases overall safety—one of the main arguments for the introduction of automated vehicles.

Frison et al. [14] came to a different conclusion. They used the trolley problem in a driving simulator study to evaluate whether people would be willing to sacrifice their own lives in favor of others. Their results suggest that people would favor a utilitarian choice even if this would mean sacrificing ones own life. Reasons for such decisions included the expectation of a system to act rationally, as well as personal consequences (e.g., not willing to possess a vehicle that could do harm to others). Fournier [13] suggest to create "command profiles" for every driver determined via a series of questions requiring ethical choices under differing and difficult driving circumstances. This profile could then be transferred to the automated vehicle.

Maurer et al. [26] recently suggested the Guardian Angel approach in which they claim that an automated vehicle should be able to override a human driver while driving manually and take over control if it detects an imminent accident. Lin [24] argues that for such a scenario there is a difference between using the trolley problem for moral decisions or looking at liability consequences. Imagine a human driver is manually driving a vehicle which is in principle capable of driving autonomously (at least SAE level 3 or above). Then the sensors of the automated vehicle foresees a fatality and actively takes over control from the human driver. By doing so it kills another person. Then the OEM might be held legally responsible for killing a person. If the vehicle would not have taken over control the driver would be legally responsible. In latter case, the OEM might only be accused for letting somebody die instead of killing somebody, which is a huge difference from a legal perspective.

Goodall [17] points out other dilemmas that arises when choosing an utilitarian choice. He gives a simple example of an automated vehicle being forced to make the choice to hit a motorcyclist with a helmet or the one without a helmet. He argues that following a utilitarian approach the car would choose to hit the rider with a helmet since he/she has better chances to survive the crash. Now the question arises, why the safety-conscious rider should be punished for his/her virtues. Finally, he argues that solutions do need to be perfect but thoughtful and defensible.

A non-intervention policy for autonomous vehicles in a trolley dilemma scenario is also discussed in the scientific discourse. Ximenes [40] provides three reasons not to program moral decisions into automated vehicles: (1) the fact that a moral machine never can explain the context in its entirety and moral decisions are heavily dependent on the these factors; (2) the impossibility to provide a metric across all nations and cultures; (3) the difficulty for humans to make complex moral judgments when the output is known.

In 2017, the Ethics Commission of the German Federal Ministry of Transport and digital Infrastructure responded to the challenges posed by automated vehicle technology with a list of 20 rules for automated vehicles [11]. Some of these concern trolley dilemmas and state, e.g., that it is not permissible to weigh human lives against each other. Neither is it permissible to program an algorithm that does so. They do, however, define a general priority of human life over animal life or inanimate objects and permit the general use of strategies to minimize the amount of casualties.

In HCI autonomous systems and automated driving was discussed recently intensely on a general level (e.g., [10], [28]), with respect to control transitions (e.g., [6]), liability (e.g., [21]), or trust (e.g., [39]). Riener et al. [33] have identified three central problem fields surrounding ethical issues in HMI for automated driving. (1) How can ethically sound decisions reached and implemented in algorithms? (2) How should ethically relevant decision making parameters be visualized to a driver / passenger? (3) Should humans or robots make these decisions? All of these discussion have in common that moral questions with respect to autonomous driving have been tackled but the general eligibility of the trolley problem have not been questioned, in our view due to a lack of a thorough philosophical discussion of such dilemmas. This paper tries to close this gap.

3 UNDERSTANDING THE TROLLEY PROBLEM

We now outline the origins of the trolley problem, present the purposes it was created for, and outline high-level solution strategies that have been pursued to tackle it.

The Trolley Problem and its Origins

Foot [12, p. 435] describes the trolley problem in her 1967 article as follows: "*The driver of a runaway tram [...] can only steer from one narrow track onto another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed.*" This short and concise description contains all the elements required for what we will continue to refer as a *trolley-like dilemma*. These are: (a) one human agent capable of making and enacting a decision, (b) a limited number of possible courses of action said individual can take, (c) at least two human individuals or groups of human individuals possibly affected by the action (d) a situation with fatal consequences for one of the involved individuals

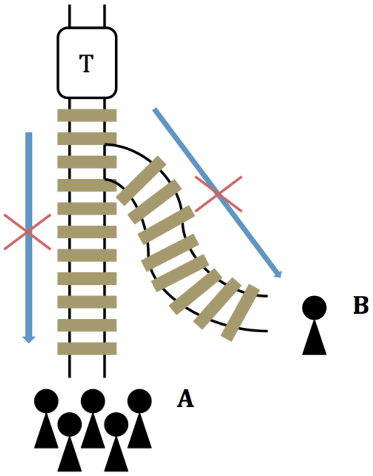

Figure 2: A simple trolley problem. Saving A means killing B and vice-versa. Both outcomes seem morally unacceptable.

(or groups) brought about by every choice possible to the human agent.

A simplified sketch of such a trolley problem can be seen in Figure 2. Note that the human agent can be one of the endangered parties in case of a self-sacrifice variant of such a dilemma. Inaction must also lead to fatal consequences. One of the reasons why the trolley-scenario is very suitable for these kinds of dilemmas is the fact that the trolley is already moving and can only move along tracks. This makes it relatively easy to create a believable scenario while still fulfilling all the requirements of a dilemma.

There are many different variations on such dilemmas. One very important general distinction is between *epistemic* and *ontological* dilemmas [4]. An epistemic dilemma refers to a conflict where the agent does not know which course of action has priority over the other, thus creating the dilemma situation. It is still possible, however, that one course of action is clearly preferable to the other one and the agent simply does not know it. An ontological dilemma, on the other hand, persist regardless of the agent's conscious state and are due to attributes of the situation itself. The latter are, therefore, also referred to as *genuine moral dilemmas* [27]. In the following, we consider the trolley problem for automated vehicles to be intended as an ontological dilemma, although we will come back to this distinction in section 4.

Trolley-like dilemmas go further back than the discussions spurred by Foot and Thomson (see e.g., Plato [32] or Sartre [35]). In order to understand the purpose of the trolley dilemma, we now take quick peek outside the realm of HCI and into classic philosophy—specifically moral theory and the debate between deontological and teleological ethics [1].

Intentions and Consequences

One of the main questions that ethics as a philosophical discipline tries to answer is, how moral judgments and normative statements can be justified. There are two basic streams or general stances regarding this question. One is deontological, also known as "consequentialism". The other is teleological, also known as "intentionalism". As their translated names imply, the idea behind the former is that any action should be judged regarding its outcomes, not the intentions behind it. The latter is the exact opposite, where the intentions matter, not the outcomes.

While these positions are more nuanced today, the classical debate often centered around which one of them was more "right" or "wrong". Right and wrong depends on whether a position can justify giving guidance to one's actions in different situations. When such situations include no moral conflicts or critical outcomes, such as, the decision to buy either chocolate or strawberry ice cream, then neither position can provide better or worse guidance than the other.

In order for a situation to be a suitable benchmark for moral actions, there are two general possibilities. It either needs have at least one outcome that clashes *or* conforms more strongly with our moral understanding than the other outcome(s). One can then argue for superiority of one position over the other by showing that one's position always mandates the moral or prohibits the immoral choice or outcome. The other possibility is to create a situation where all possible courses of action clash with our moral sense in some way. Such *lemmas* can then be used to show that they only hold when one position is applied but not the other.

The Search for a Normative Basis

Roughly speaking, ethics concerns itself with the search for morality. It is the scientific discipline that examines morality and morals, whereas morals are the principles that tell the individuals in a society what they ought and ought not do. As human beings, we have stronger emotional responses to some issues than we do towards others. For example, the average human being is much more inclined to refuse to harm another human being than she/he is to refuse an invitation to a wedding. This is, in part, because one is a matter of preference, the other one of morality. The laws that guide a society are supposed to reflect the sense of morality of said society, i.e., the moral norms that everyone is supposed to follow. The question is, then, what exactly these moral norms are and how they can be discovered or verified.

Since it seems unsatisfactory to assume that these norms are completely arbitrary and since we seem to have such common emotional responses to certain situations, one common assumption was to postulate general or *universal* moral principles. While some might differ between individuals or societies, these universal norms would apply to everyone and constitute the basis for a common and non-arbitrary moral system (and legal system as a consequence). Defending such a "moral realism" [18] means having to find these universal norms or at least plausibly suggest their existence.

Normative statements are different from descriptive statements in that they provide guidance to act. So if one were to "discover" an objective norm, then that norm would be able to tell an individual what she/he should do in a situation the norm applies to in order to act morally sound. Thus, the way to dispute the status of a norm as such would be to provide an example of a situation in which it provides guidance that goes against to our intuitive understanding of what is morally right.

Trolley-like dilemmas are a commonly used way to suggest just that. They do so by taking the one principle that seems like the most likely candidate for a universal norm: "*Thou shalt not kill!*". By providing an example of a realistic situation in which an individual kills another human being, the trolley argument can be used to dispute the status of "*Thou shalt not kill!*" as a universal norm and, thereby, the existence of universal norms as a whole.

Once again, the proposed situation is unsolvable. It intends to highlight a fundamental issue in the position that is being attacked. Which action the individual takes to arrive at the inevitable outcome is not relevant, even moreso than before, as the individual and its motivations are not under investigation in this context. While the specific nuances of these discussions are not directly relevant to the topic at hand, they serve to highlight the nature of such a problem. It is *designed to be unsolvable* by those challenging a certain position. These positions concern meta-issues; the choices available to the individual within the dilemma are a proxy for the overall argument and not under investigation themselves.

4 "SOLVING" THE TROLLEY PROBLEM

In this section, we will first outline different strategies that can or have been pursued to try and solve trolley-like dilemmas. Based on a treatise of the most important aspects in philosophical literature surrounding the topic, we provide a general summary of ways to tackle such problems. We will then relate these solutions to the trolley problem for automated vehicles, to see whether these would be appropriate within that context.

Solving an Ethical Dilemma

Just because a dilemma is designed to be unsolveable does not mean that there have not been attempts to solve it. There are indeed ways that have been proposed to resolve or circumvent dilemma situations. One strategy is to deny the existence or possibility of genuine moral dilemmas (e.g., Zimmerman [41]). This can be done by showing that the construction of a dilemma leads to logical inconsistencies given a principle that is more desirable to hold than the possibility of genuine dilemmas.

One such principle is that of *deontic consistency* (see [27]), which essentially states that when a certain action ought to be done, then it can not also be forbidden at the same time. Since a dilemma is characterized by all possible actions available to the agent being forbidden, this creates an inconsistency, leading to a possible argument against the possibility of genuine moral dilemmas[3]. This would then mean that a dilemma would only appear as such to the observer due to incomplete information. Thus, the only possible dilemmas are *epistemological* ones, which only constitute a problem of incomplete information, not one of fundamental nature.

Another possible strategy is to deny the symmetry [36] and allow for different levels of priority between the courses of action that is often assumed when creating an ethical dilemma. The mere fact that all outcomes lead to morally wrong actions or outcomes, does not automatically mean that all of them are equally immoral, e.g., the act of stealing medicine in order to save someone's life. In this case, violating the norm not to steal would be overruled by the norm to save the lives of others. This would not change the validity of the norm to not steal in isolation, it would simply be of a lower priority than the norm to save others, which seems consistent with moral intuition.

It is different, however, if both outcomes violate the same norm, as in this case, it is not possible to state that one norm has higher priority than itself. What is still possible, however, is to state that one course of action causes more or more severe violations of the same rule than the other. This would be possible with Foot's version of the trolley problem, where the choice lies between one or five dead workers [12]. This can be translated into a choice between one or five violations of the same norm, at which point the "correct" choice becomes clear.

In case that both choices *are* perfectly symmetrical, e.g., when both choices lead to exactly one dead individual, it can still be argued that it is wrong to assume that the agent does wrong, no matter the choice she/he makes. McConnell [27] expresses this as follows: "*Such a move need not be ad-hoc, since in many cases it is quite natural. If an agent can afford to make a meaningful contribution to only one charity, the fact that there are several worthwhile candidates does not prompt many to say that the agent will fail morally no matter what*

[3] Note that the full argument is more complicated than this and requires assumption of a second commonly held deontic principle. The basic idea, however, remains the same.

he does. Nearly all of us think that he should give to one or the other of the worthy candidates. Similarly, if two people are drowning and an agent is situated so that she can save either of the two but only one, few say that she is doing wrong no matter which person she saves." Such ways of argumentation are very compatible with utilitarian viewpoints, which allow weighing different options—including the preservation and loss of life—depending on their benefit to the overall good.

A third way to respond to a trolley-like dilemma is to assume a teleological stance and consider the intentions of the agent when deciding on their course of action (for an example, see e.g., Kagan [22]). This is one of the easier ways to solve a trolley-like dilemma, as they are usually constructed as such that the *consequences* of all possible courses of action seem morally unacceptable, thus causing one to assume consequentialistic viewpoint from the start. But if one considers intentions to be the primary factor behind the morality of an action, then the teleologist could, just like the consequentialist, argue for saving the group of five instead of only one. But instead of justifying it via the outcome, she/he can argue that their intention was to save a group of five people rather than killing the other one. Thus, the intentionalist can claim that the consequentialist's perspective does not matter, as both action's outcomes lead to a violation of the norm to not kill in their consequences, while the intentions can be interpreted in a positive way, thus not causing *intentional* norm violations.

Note, however, that this reasoning can be applied to both cases, as intending to save one or five individuals are both usually considered to be good intentions. Especially in symmetric cases (e.g., saving one individual vs. saving a different one), the intentionalist perspective does not provide instructions on what to do in the dilemma situation, as the agent ought to save both—which he is unable to do.

Finally, it is possible to distinguish between *self-imposed* dilemmas and those *imposed on an agent by the world* [9, 27]. A self-imposed dilemma, as the name suggests, exists due to the agent's actions prior to the dilemma situation. It can then be argued that such a dilemma is not genuine in the sense that the agent already acted wrong in bringing about the dilemma situation.

Thus, there was a point prior to the dilemma at which the agent had a choice between right and wrong and an adequate moral theory could have told the agent what to do. But since the agent chose poorly, the dilemma came about. Resolving a dilemma in this sense means to identify this action. If it is not contained in the description of the dilemma itself (as is mostly the case with trolley-like dilemmas), it can be argued that the dilemma itself is an *incomplete description*, as it contains only the consequences of the action that really matters in the moral sense.

Dilemma Solutions in Automated Driving

Let us now try to apply these general solution strategies to the context of automated driving. Remember that, in the end, the vehicle will have to be able to drive on the road making decisions on where to drive and where not to, thus requiring a means to decide what do to in any given situation.

If we assume the position that genuine ontological dilemmas are impossible, then the trolley problem for automated vehicles can also not be a genuine dilemma. It, therefore, must be of epistemological nature, meaning that any such situation can be resolved when sufficient information is available. This in itself is not sufficient to tell the vehicle how to resolve the situation of whether to kill one individual over another.

What it does seem to tell us is that, once we learn enough about the situation and the actors within it, we will arrive at the conclusion of which life should be saved over the other. This suggests a utilitarian solution, where a method to calculate the value of any human life does exist but it simply has not been found yet. Providing such a solution requires a complete assessment of any such driving situation that does not miss even a single potentially relevant detail. This is theoretically feasible, given appropriately sophisticated sensor technology, detection algorithms, etc. However, this is only one part of the solution and the only one that technical progress can provide.

In order to then make the situation assessment and calculate the "right" course of action, a universally valid utilitarian calculus is needed that assigns values to individuals, with modifiers based on attributes (age, occupation, etc.). The attainability of such a calculus is disputed (see e.g., Moore [29] and Lin [24]) and the assignment of such *objective* values in a manner that is intuitively more agreeable than any other solution that kills one or the other, is doubtful.

The second strategy leads to a similar outcome, as one would then assume that the available options in a dilemma situation are both undesirable but not symmetrical. This means that one is even more undesirable than the other, making the least undesirable option the "correct" choice. Once again, the vehicle must then make an assessment of these and calculate this correct choice. While this does not necessarily entail utilitarianism, it is one of the few existing ways to link morality to a numeric calculus, which is what the machine will ultimately require to make its decision. Whichever conclusion the vehicle comes to in the end, the solution will result in it killing one individual or the group of individuals but not the other.

The third strategy is, once again, not very different from the first and second one when applied to the automated driving context. By taking away the focus from the outcomes and looking at the intentions instead, one can try and justify the vehicle's actions based on its intentions when making the choice. This choice will, naturally, have to depend on a calculus weighing the different options available, thus providing no further input about what that calculus might look like in the end, though a utilitarian one would seem likely once more.

Unlike the first two strategies, this one can serve a very interesting function on the social level, as the discussion goes away from the consequences, which are always fatal and, thus, always disagreeable. Intentions need not always be disagreeable, even in a dilemma situation. Designing an artificial intelligence that saves lives sounds much better than one that takes them—even if one entails the other.

Ultimately, it is still a matter of perspective: once the vehicle has made its decision, someone will die and someone will live. Some will sleep better that night, thinking that the vehicle had good intentions, others will not, thinking that it decided to take a life.

This leaves us with the final strategy. If we consider all ethical dilemmas to be self imposed, then the automated vehicle trolley problems must be self-imposed as well. This way, we would not be confronted with a genuine dilemma and the strategy to solve it would consist of identifying the cause that led to the dilemma situation in the first place.

A standard trolley description, as it is usually found in literature, is a poor fit for this purpose most of the time. It contains no information regarding who put the trolley in motion in the first place, why there are several workers on the tracks, supposedly under the assumption that no trolley would drive on them on that day, why there is an apparently freely accessible lever to change the course of the trolley, and so on. In a way, this is a different perspective on the trolley problem's status as an epistemological dilemma. The solution consist on shifting the blame or responsibilty to whoever was responsible in the first place. Within the scope of a typical trolley problem description, this is not possible, as such information is not provided, nor are the means to acquire this information. But even if this information were available, such a solution would not say anything regarding right and wrong in the dilemma situation itself.

Any action that causally led to the dilemma situation *prior* to the actual dilemma was morally wrong. At that point of the trolley dilemma, these actions have already concluded and are, thus, out of scope in a certain sense. Certainly, one could say that the agent is either morally wrong or innocent no matter which course of action they choose, depending on whether she/he or someone else caused the dilemma situation in the first place. Regardless of this, however, the individual choice within the dilemma does not matter from this perspective. The agent can either do no wrong, in case he/she was not responsible for the dilemma, or can do only wrong, in case he/she was responsible.

In the end, neither strategy can tell the vehicle more than kill either one or the other involved party. The reasoning might certainly change but the outcome will not. This is because a trolley-like dilemma is specifically constructed in an idealized way, often relying on incomplete information, in order to bring about a situation in which the *individual decision no longer matters.* It is only natural then, that strategies to solve such a problem rarely involve instructions to act for the agent within the dilemma.

5 DISCUSSION

It would seem that the trolley problem, by itself, will not bring us any closer to an answer on what the vehicle should do. This is not specific to the trolley problem but a general feature of such dilemmas. As McConnell [27] explains : "*It will [be] tempting for supporters of dilemmas to say to opponents, 'If this is not a real dilemma, then tell me what the agent ought to do and why?' It is obvious, however, that attempting to answer such questions is fruitless, and for at least two reasons. First, any answer given to the question is likely to be controversial, certainly not always convincing. And second, this is a game that will never end; example after example can be produced. The more appropriate response on the part of foes of dilemmas is to deny that they need to answer the question.*" But it would be wrong to assume from this that the trolley problem is completely without merit within the context of automated driving. It still is a powerful tool to raise the issue of rule conflicts and the overall discussion around permissible levels of autonomy in autonomous agents.

At the same time, however, it is a poor benchmark for the success of such autonomous agents' decision making. Confronting automated vehicle technology with the trolley problem and then judging said technology on their decision taken within the dilemma is akin to posing the same problem to an accused human individual at court, and then convicting them anyway regardless of their response because there was never a right choice to begin with. Solving a trolley dilemma amounts to showing that the problem does not apply or only seems to apply because of an incomplete or incorrect description of the situation. By positioning the automated vehicle inside of such a dilemma with the only avenues of answer being to kill either x or y, we take away all possibilities to sensibly respond to the problem in a constructive manner.

The Anthropocentric Fallacy

One aspect that makes the trolley problem so compelling within the automated driving context lies with the assumed intentionality behind any given action pursued within the dilemma. It is not simply a situation in which a fatal outcome inevitable. There is a machine, an artificially created being, that *decides.*

It seems that there is a, somewhat inconsistent, anthropocentric stance towards automation technology as far as the trolley problem is concerned. It is inconsistent because, on the one hand, we seem to implicitly assume that an autonomous agent is able to make a conscious decision just like a human would do. This decision can include malice, carelessness, and other aspects of human intentionality that can make such acts morally condemnable.

On the other hand, when a human is confronted with a dilemma situation, their actions and intentions are being investigated relative to their abilities and decision making processes. No judge will ever ask a human individual what their utilitarian calculus was that made them decide to kill an 80 year old woman over a group of children. Rather, they will be asked whether they possessed a valid driving licence, were under the influence of alcohol, strong medication, or other mind-altering substances, whether they suffered from lack of sleep, had any physical impairments, and so on. The trolley problem is mainly a problem of consequences, whereas most judicial systems do also take intentions into account.

When judging an autonomous agent purely on the consequences of their actions in a pre-constructed dilemma that sets them up for failure, then we do not grant them the courtesy they should have by virtue of being a moral agent. That is not to say that they are exempt from such judgment but dilemmas are as unsolveable for humans as they are for autonomous agents. If looking only at the consequences does not work for humans, it should not be expected to work for autonomous agents either. Rather, their behaviour should be judged based on their *individual* capabilities and decision making processes. This is what Holstein et al. [20] refer to when they speak of a certain 'intrinsic unfairness' of the trolley problem. The standard that is being applied when judging the consequences of their actions is not maintained when judging their intentions. It is at this point where the trolley problem can be very helpful in guiding the way forward in decision making for automated vehicles.

What Really Matters

Nyholm [31] states: "*when it comes to the real world issue of the introduction of self-driving cars into real world traffic, we cannot do what those who discuss the trolley problem do. We cannot stipulate away all considerations having to do with moral and legal responsibility. We must instead treat the question of how self-driving cars ought to be pre-programmed as partly being a matter of what people can be held morally and legally responsible for.*" Goodall [17] correctly identifies the morally relevant decisions for vehicles in dilemma situations as "*prospective decisions or contingency-planning*", as opposed to in-the-moment decisions.

When looking at decision making in automated vehicles, we should use the trolley problem not as a collection of two possible outcomes but rather consider the trolley problem itself as one possible outcome among several. The trolley problem is the one that needs to be avoided. Thus, the decision the vehicle would take within the dilemma is not the benchmark for its success, but rather what it does or can do to avoid ending up in the dilemma situation.

Goodall [16] proposes risk management strategies as a possible means to determine the optimal course of action for automated vehicles. Holstein et al. [20] propose a closer analysis of the vehicle's decision making process and basing the assessment on the reliability of the software that mediates between the vehicle's components (e.g., processing and interpretation of sensor data). The report by the German Ethics Commission [11] also specifically highlights the post-hoc character of accident assessments and suggests to establish an independent institution for the collection and analysis of accident data with automated vehicles.

Based on our analysis, we can conclude that a moral judgement of a decision taken within a trolley problem is of little use. Instead, efforts should be directed towards avoiding such situations. In order to do so, additional information is needed, although it is not yet entirely clear what the specific elements of that equation are. Whichever particular form it will take in the end, a solution to the trolley problem for automated vehicles entails knowing the agents involved in any given situation and their capabilities. This includes, but is not limited to, the vehicle's own capabilities influenced by the circumstances (e.g., weather, type of road, road condition), risk assessments based on possible courses of action in a given situation (including parameters such as speed or likelihood of other traffic participants and their vulnerability), likelihood of technical failure, software errors, and sensor inaccuracies (including the capabilities and constraints of certain sensors and their interplay with decision-making algorithms). This is where HCI can contribute.

Instead of asking the vehicle who it would kill in case of an ethical dilemma, it seems much more sensible to ask whether it correctly switches between radar and lidar depending on weather conditions, whether it can compensate for radar interference, or whether it's operating system is up do date. We argue that much more emphasis has to be put on aspects such as the interrelation between sensor quality, avoidance algorithms, decision algorithms, and access to vehicle-to-vehicle infrastructure. Note that none of these aspects are ethical in nature and yet appear to be the way forward to improve decision making in vehicles from both the technical *and* ethical perspective.

The reason for this is that, as we have shown, programming to specifically solve the trolley problem does not work for conceptual reasons, thus neither regulatory nor implementation activities should use it as the sole benchmark. With respect to interface design, we see a design space which addresses the need for information and transparency of automated vehicle decisions for the driver, passenger, or operator. For example, if an automated vehicle makes the decision to drive slower because of sensor inaccuracy or environmental factors this should be visualized. Thereby solutions, that have been proposed to present information to a driver/passenger as for example uncertainty displays implemented in Augmented Reality (see e.g., [23]) may be utilized. To conclude we see a shift from designing for the trolley problem towards design to communicate decisions the vehicle makes in order to avoid potentially hazardous situations—including trolley-like ones.

6 CONCLUSIONS

Trolley dilemmas seem very important within the context of automated driving, as they highlight a potential authority of machines over human life that society is uncomfortable with. In this way, the trolley problem serves a very important function in society in that it allows us to reflect on potential consequences of the ever increasing degree of automation and the explore the boundaries of what it means to be a conscious human being.

As benchmarks for the success of an automation algorithm's decision making capabilities, however, they are unfit. We showed this by outlining the origins of the trolley problem, solution strategies, and how these relate to automated driving. None of the solutions worked in the way one would expect to solve the trolley problem for automated vehicles. This is because, ultimately, what we are interested in is an answer to the question "*What should the vehicle do?*" But none of the described solution strategies ever attempted to do so, as that would have been fruitless.

Trolley-like dilemmas are constructed in a way so that the decision on what to do does not matter, so that the meta-discussion (be it about universal norms or deontologism versus teleologism) can be triggered. We concluded by proposing to consider the trolley problem as not a decision the machine must be able to resolve but one that it must be able to anticipate and avoid. Thus, the correct way to address the trolley problem is to simply continue existing work to improve decision making based on available environmental information, only with a specific focus on defining and identifying different situation types, avoidance of certain types, and visualization of decision making in vehicles.

ACKNOWLEDGMENTS

The financial support by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit) and the Austrian Research Promotion Agency (FFG) under grant number 865110 (Project: Digibus Austria) within the program "Mobility of the Future" are gratefully acknowledged.

REFERENCES

[1] Larry Alexander and Michael Moore. 2016. Deontological Ethics. In *The Stanford Encyclopedia of Philosophy* (winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford, CA, USA.

[2] BBC Radio 4 2014. Right & Wrong: The Trolley Problem. Retrieved Sep 19, 2018 from https://www.youtube.com/watch?v=bOpf6KcWYyw

[3] Hanna Bellem, Barbara Thiel, Michael Schrauf, and Josef F. Krems. 2018. Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transportation Research Part F: Traffic Psychology and Behaviour* 55 (2018), 90 – 100. https://doi.org/10.1016/j.trf.2018.02.036

[4] Simon Blackburn. 1996. Dilemmas: Dithering, Plumping, and Grief. In *Moral Dilemmas and Moral Theory*, H. E. Mason (Ed.). Oxford University Press, New York, NY, USA, 127.

[5] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[6] Shadan Sadeghian Borojeni, Alexander Meschtscherjakov, Alexander G. Mirnig, Susanne Boll, Frederik Naujoks, Ioannis Politis, and Ignacio Alverez. 2017. Control Transition Workshop: Handover and Takeover Procedures in Highly Automated Driving. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct (AutomotiveUI '17)*. ACM, New York, NY, USA, 39–46. https://doi.org/10.1145/3131726.3131732

[7] Barry Brown, Alexandra Weilenmann, Donald McMillan, and Airi Lampinen. 2016. Five Provocations for Ethical HCI Research. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 852–863. https://doi.org/10.1145/2858036.2858313

[8] Stephen M. Casner, Edwin L. Hutchins, and Don Norman. 2016. The Challenges of Partially Automated Driving. *Commun. ACM* 59, 5 (April 2016), 70–77. https://doi.org/10.1145/2830565

[9] Alan Donagan. 1993. Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy. *Ethics* 104, 1 (1993), 7–21. https://doi.org/10.1086/293573

[10] Michael Feary, Célia Martinie, Philippe Palanque, and Manfred Tscheligi. 2016. Multiple Views on Safety-Critical Automation: Aircrafts, Autonomous Vehicles, Air Traffic Management and Satellite Ground Segments Perspectives. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1069–1072. https://doi.org/10.1145/2851581.2886430

[11] Federal Ministry of Transport and Digital Infrastructure 2017. Ethics Commission for Automated and Connected Driving: June 2017 Report [DE: Ethik-Kommission Automatisiertes und Vernetztes Fahren: Bericht Juni 2017]. Retrieved Dec 27, 2018 from https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf

[12] Philippa Foot. 1967. The Problem of Abortion and the Doctrine of the Double Effect*. *Virtues and Vices: and other essays in moral philosophy* 5 (1967), 5–15. https://doi.org/10.1093/0199252866.003.0002

[13] Tom Fournier. 2016. Will my next car be a libertarian or a utilitarian?: Who will decide? *IEEE Technology and Society Magazine* 35, 2 (2016), 40–45.

[14] Anna-Katharina Frison, Philipp Wintersberger, and Andreas Riener. 2016. First Person Trolley Problem: Evaluation of Drivers' Ethical Decisions in a Driving Simulator. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16 Adjunct)*. ACM, New York, NY, USA, 117–122. https://doi.org/10.1145/3004323.3004336

[15] Noah J. Goodall. 2014. Ethical Decision Making During Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board* 2424 (2014), 58–65. https://doi.org/10.3141/2424-07

[16] Noah J. Goodall. 2016. Away from Trolley Problems and Toward Risk Management. *Applied Artificial Intelligence* 30, 8 (2016), 810–821. https://doi.org/10.1080/08839514.2016.1229922

[17] Noah J Goodall. 2016. Can you program ethics into a self-driving car? *IEEE Spectrum* 53, 6 (June 2016), 28–58. https://doi.org/10.1109/MSPEC.2016.7473149

[18] Gilbert Harman. 1977. *The Nature of Morality: An Introduction to Ethics.* Oxford University Press, New York, NY, USA.

[19] Franziska Hartwich, Matthias Beggiato, and Josef F. Krems. 2018. Driving comfort, enjoyment and acceptance of automated driving - effects of driver's age and driving style familiarity. *Ergonomics* 61, 8 (2018), 1017–1032. https://doi.org/10.1080/00140139.2018.1441448

[20] Tobias Holstein and Gordana Dodig-Crnkovic. 2018. Avoiding the Intrinsic Unfairness of the Trolley Problem. In *Proceedings of the International Workshop on Software Fairness (FairWare '18).* ACM, New York, NY, USA, 32–37. https://doi.org/10.1145/3194770.3194772

[21] Michael Inners and Andrew L. Kun. 2017. Beyond Liability: Legal Issues of Human-Machine Interaction for Automated Vehicles. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '17).* ACM, New York, NY, USA, 245–253. https://doi.org/10.1145/3122986.3123005

[22] Shelly Kagan. 1989. *The Limits of Morality.* Oxford University Press, New York, NY, USA.

[23] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2018. Augmented Reality Displays for Communicating Uncertainty Information in Automated Driving. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18).* ACM, New York, NY, USA, 164–175. https://doi.org/10.1145/3239060.3239074

[24] Patrick Lin. 2016. *Why Ethics Matters for Autonomous Cars.* Springer Berlin Heidelberg, Berlin, Heidelberg, 69–85. https://doi.org/10.1007/978-3-662-48847-8_4

[25] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15).* ACM, New York, NY, USA, 117–124. https://doi.org/10.1145/2696454.2696458

[26] Steffen Maurer, Rainer Erbach, Issam Kraiem, Susanne Kuhnert, Petra Grimm, and Enrico Rukzio. 2018. Designing a Guardian Angel: Giving an Automated Vehicle the Possibility to Override Its Driver. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18).* ACM, New York, NY, USA, 341–350. https://doi.org/10.1145/3239060.3239078

[27] Terrance McConnell. 2018. Moral Dilemmas. In *The Stanford Encyclopedia of Philosophy* (fall 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford, CA, USA.

[28] Alexander Meschtscherjakov, Manfred Tscheligi, Bastian Pfleging, Shadan Sadeghian Borojeni, Wendy Ju, Philippe Palanque, Andreas Riener, Bilge Mutlu, and Andrew L. Kun. 2018. Interacting with Autonomous Vehicles: Learning from Other Domains. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18).* ACM, New York, NY, USA, Article W30, 8 pages. https://doi.org/10.1145/3170427.3170614

[29] George Edward Moore. 1903. *Principia Ethica.* Cambridge University Press, Cambridge, UK.

[30] National Highway Traffic Safety Administration (NHTSA) 2017. Automated Driving Systems 2.0: A Vision for Safety. Retrieved Dec 27, 2018 from https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf

[31] Sven Nyholm and Jilles Smids. 2016. The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice* 19, 5 (01 Nov 2016), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

[32] Plato. 1930. The Republic, trans, Paul Shorey. In *The Collected Dialogues of Plato*, Edith Hamilton and Huntington Cairns (Eds.). Princeton University Press, Princeton, NJ, USA.

[33] Andreas Riener, Myounghoon Philart Jeon, Ignacio Alvarez, Bastian Pfleging, Alexander Mirnig, Manfred Tscheligi, and Lewis Chuang. 2016. 1st Workshop on Ethically Inspired User Interfaces for Automated Driving. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16 Adjunct).* ACM, New York, NY, USA, 217–220. https://doi.org/10.1145/3004323.3005687

[34] SAE International 2018. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard J3016-2018. Retrieved Dec 27, 2018 from https://www.sae.org/standards/content/j3016_201806/

[35] Jean-Paul Sartre. 2007. *Existentialism is a Humanism.* Yale University Press, New Haven, CT, USA. https://books.google.at/books?id=IfNqvQXbuk8C

[36] Walter Sinnott-Armstrong. 1984. 'Ought' Conversationally Implies 'Can'. *Philosophical Review* 93, 2 (1984), 249–261.

[37] The Guardian 2016. The trolley problem: would you kill one person to save many others? Retrieved Dec 27, 2018 from https://www.theguardian.com/science/head-quarters/2016/dec/12/the-trolley-problem-would-you-kill-one-person-to-save-many-others

[38] Judith Jarvis Thomson. 1976. Killing, Letting Die, and the Trolley Problem. *The Monist* 59, 2 (1976), 204–217.

[39] Philipp Wintersberger, Brittany E. Noah, Johannes Kraus, Roderick McCall, Alexander G. Mirnig, Alexander Kunze, Shailie Thakkar, and Bruce N. Walker. 2018. Second Workshop on Trust in the Age of Automated Driving. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18).* ACM, New York, NY, USA, 56–64. https://doi.org/10.1145/3239092.3239099

[40] Bianca Helena Ximenes. 2018. Non-intervention Policy for Autonomous Cars in a Trolley Dilemma Scenario. *AI Matters* 4, 2 (July 2018), 33–36. https://doi.org/10.1145/3236644.3236654

[41] Michael J. Zimmerman. 2000. The Concept of Moral Obligation. *Philosophical and Phenomenological Research* 60, 1 (2000), 242–244.