Technische Universität München
TUM School of Social Sciences and Technology

The risk ethics of autonomous vehicles

Sebastian Krügel

Vollständiger Abdruck der von der TUM School of Social Sciences and Technology der Technischen Universität München zur Erlangung eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Jürgen Pfeffer

Prüfende der Dissertation:

1. Prof. Dr. Matthias Uhl
2. Prof. Dr. Christoph Lütge

Die Dissertation wurde am 05.07.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Social Sciences and Technology am 01.11.2024 angenommen.

Abstract

Autonomous vehicles (AVs) continuously have to weigh up risks between different courses of action, all of which might also impact other road users. For example, a minimal readjustment of AVs' lateral lane positioning may reduce the probability of collision for one road user while increasing the probability of collision for another. In this dissertation, I first elaborate that much of the ethics of AVs is primarily concerned with the allocation of casualties in unavoidable collision scenarios, while vehicle engineering is primarily concerned with collision avoidance of AVs. I then point out that risks in road traffic relate to both dimensions and argue that the ethics of AVs should mainly be concerned with the desired risk management of AVs. However, standard moral theories swiftly reach their limits when assessing risks. This dissertation contributes to this debate through the lens of data-driven ethics. In two research projects, I first develop trolley-like dilemma situations further by explicitly incorporating risks into the scenarios to be examined and then survey how people trade off the probability and severity of collisions. It turns out that both dimensions are important for the study participants of both research projects. A focus on pure collision avoidance would therefore not be in line with the moral preferences of the participants. In a third research project, I additionally investigate the importance of appropriate behavior by AVs in the event of a collision. This is relevant because the technology in AVs is often not yet designed to reliably detect accidents and current regulations on AVs do not require such detection. The study participants of the third research project expressed a strong preference that AVs should be able to detect collisions and behave appropriately. The participants additionally indicated a clear willingness to pay for the necessary technology.

Deutsche Zusammenfassung

Autonome Fahrzeuge (AVs) müssen ständig Risiken zwischen verschiedenen Handlungsalternativen abwägen, die auch Auswirkungen auf andere Verkehrsteilnehmer haben. So kann zum Beispiel eine minimale Adjustierung der lateralen Fahrposition eine Verringerung der Kollisionswahrscheinlichkeit für einen Verkehrsteilnehmer nach sich ziehen, während sie für einen anderen Verkehrsteilnehmer erhöht wird. In der vorliegenden Dissertation arbeite ich zunächst heraus, dass sich ein Großteil der Ethik der AVs hauptsächlich mit der Verteilung von Personenschäden bei unvermeidlichen Kollisionsszenarien beschäftigt, während sich die Fahrzeugtechnik vornehmlich mit der Kollisionsvermeidung von AVs auseinandersetzt. Anschließend mache ich deutlich, dass Risiken im Straßenverkehr von beiden Dimensionen abhängen und argumentiere, dass sich die Ethik vor allem mit dem gewünschten Risikomanagement von AVs beschäftigen sollte. Jedoch stoßen gängige Moraltheorien bei der Bewertung von Risiken schnell an ihre Grenzen. Die vorliegende Dissertation trägt zu dieser Debatte aus der Perspektive einer datengetriebenen Ethik bei. In zwei Forschungsprojekten entwickle ich zunächst Trolley-ähnliche Dilemmasituationen weiter, indem ich explizit Risiken in die zu untersuchenden Szenarien einbinde und erhebe anschließend, wie Menschen Kollisionswahrscheinlichkeit und -schwere gegeneinander abwägen. Es zeigt sich, dass für die Studienteilnehmer beide Dimensionen wichtig sind. Ein Fokus auf die reine Kollisionsvermeidung wäre damit nicht im Einklang mit den moralischen Präferenzen der Teilnehmer. In einem dritten Forschungsprojekt untersuche ich darüber hinaus, welchen Stellenwert ein angemessenes Verhalten von AVs im Falle eines Unfalls hat. Dies ist relevant, da die Technologie in AVs bislang häufig nicht dafür ausgelegt ist, Unfälle zuverlässig zu detektieren und aktuelle Regularien zu AVs dies auch nicht vorschreiben. Die Studienteilnehmer äußerten eine starke Präferenz, dass AVs fähig sein sollten, Kollisionen zu erkennen und sich angemessen zu verhalten. Sie brachten außerdem eine deutliche Zahlungsbereitschaft für die dafür notwendige Technik zum Ausdruck.

Acknowledgments

First and foremost, I would like to thank my doctoral supervisor, Professor Matthias Uhl, for his constant support, his constructive feedback and the pleasant collaboration throughout my doctoral studies. His profound knowledge, his ability of abstraction and his creativity are extraordinary and have helped me a lot with my research endeavors. I always encountered open doors and received patient and valuable advice.

In addition, I am grateful to Prof. Christoph Lütge for his unconditional willingness to be my second supervisor and reviewer of my doctoral thesis. I learned a lot from him in the many seminars and workshops.

I would also like to thank Prof. Jürgen Pfeffer for his spontaneous readiness to chair the examination committee.

Special thanks go to all my current and former colleagues from TUM and THI for creating a pleasant, friendly and productive research environment. Many of them have contributed to this work through discussions and insightful comments.

Finally, I thank my beloved wife Ramona and my two lovely children Lennard and Matteo for their understanding for the many extra hours sitting at my desk and for their mental support to finish this thesis. Thank you so much, my dears!

Contents

1 Introduction

In August 2023, *Cruise* and *Waymo* received full approval to transport paying passengers throughout the city of San Francisco 24 hours a day in autonomous vehicles (AVs) without a safety driver. Two months later, *Cruise's* license was initially revoked after one of its self-driving cabs ran over a pedestrian and dragged her several meters (Cano, 2024). *Waymo* cabs have caught attention for unusual behavior as well. One *Waymo* cab, for instance, drove down the wrong side of the road for almost half a minute to pass unicyclists and scooters (Connatser, 2024). Another *Waymo* cab caused a traffic jam during San Francisco's morning rush hour because it stopped in the middle of a key intersection (Bote, 2023).

These examples show that AVs are already used extensively in road traffic without safety drivers and that they can cause detrimental events in road traffic. They pose traffic risks through unusual behavior, cause obstructions in road traffic or are even involved in serious accidents. All these examples illustrate that the driving behavior of AVs has serious ethical implications. In recent years, the academic literature has addressed ethical issues related to AVs as well. This debate is dominated by moral dilemmas in unavoidable accident situations. On the one hand, AVs are expected to increase road safety overall (Lütge, 2017), but on the other hand, AVs will not be able to eliminate accident risks entirely (Goodall, 2016a), as the above examples show.

This doctoral thesis contributes to this debate. It is argued that ethical issues do not only become relevant in the face of unavoidable accident situations, but that everyday traffic situations have ethical implications due to the distribution of traffic risks among road users. It is pointed out that the ethical and engineering literature on AVs focus on different objectives in road traffic and thus partly talk past each other. This thesis emphasizes the need for a risk ethics approach that combines both objectives.

Methodologically, the doctoral thesis follows a participatory approach that elicits the moral preferences of the general public in order to investigate citizens' attitudes towards the distribution of risks in road traffic. This approach is considered appropriate in this thesis because the citizens themselves are those who participate in road traffic and should therefore have a say in determining which distributions of risks are considered acceptable.

In the following chapter, I first present the current state of the empirical ethics of AVs, before presenting common points of criticism of the research approach in Chapter 3. In Chapter 4, I summarize the research projects conducted in this doctoral thesis and then discuss the results against the background of the current academic debate in Chapter 5. Chapter 6 highlights the most important findings of the thesis and provides an outlook for future research.

2 The current state of the empirical ethics of AVs

The empirical ethics of AVs mainly rely on variants of the *trolley problem* (Foot, 1967; Greene, 2013). The starting point is usually the so-called *switch dilemma*, in which a driverless trolley is hurtling out of control towards five people and these five people can only be saved if the trolley is diverted onto another track, thereby killing one other person. The analogy to possible situations involving AVs in road traffic appears to be obvious. Although such dilemmas are presumably very unlikely, it seems inevitable that they will occur once AVs become widespread (Greene, 2016). How should AVs solve such dilemmas? What does the public expect here? And on what factors do these expectations or moral intuitions of the public depend?

With their empirical study, Bonnefon, Shariff and Rahwan (2016) initiated the "data-driven study of driverless car ethics." The subsequent most ambitious and comprehensive attempt to determine the moral attitudes of the international public in ethically relevant situations of AVs is undeniably the "Moral Machine experiment" (Awad et al., 2018). In this chapter, I will present these two studies as exemplary cases due to their importance and prominence in the literature on empirical ethics of AVs. There are of course many other studies that examine the moral intuitions of the population in dilemmatic situations involving AVs (see, e.g., Faulhaber et al., 2019; Frank et al., 2019; Huang, Greene and Bazerman, 2019; Meder et al., 2019; Bigman and Gray, 2020; Morita and Managi, 2020). However, their study design and methods are essentially the same. It is therefore sufficient for the critical discussion of the methodological approach in the subsequent chapter to present the two main studies here.

Bonnefon, Shariff and Rahwan (2016) conducted six studies with several modified variants of the *switch dilemma*. The studies were designed as online surveys and participants were recruited through Amazon Mechanical Turk (MTurk). In all surveys, the moral dilemma was presented to participants in the form of a vignette – i.e., a short written description of a hypothetical situation. All situations involved an AV driving on a

main road with one (or two) passenger(s). Suddenly, pedestrians appear in front of the AV (in several variants between 1 and 100 pedestrians). These pedestrians will be killed if the AV does not take evasive action. If the AV swerves, the pedestrians will be saved, but the passenger(s) in the AV will die as a result of the evasive maneuver. In some situations, participants were asked to imagine that they or a family member were sitting in the AV. In others, the passengers of the AV were anonymous persons.

Overall, the participants considered it more moral for the AV to sacrifice its own passengers if more pedestrians could be saved as a result. The more pedestrians that could be saved, the stronger the moral approval to sacrifice the passengers. This was true even if the participants themselves or a family member were passengers in the AV, although moral approval was much less pronounced here overall. However, the participants reported that they were much less willing to buy AVs in which they could be sacrificed as passengers. This is the so-called “social dilemma of AVs.” People find utilitarian AVs morally better and prefer others to use them. However, they themselves would not want to buy and ride utilitarian AVs, but would rather prefer AVs that protect their passengers at all costs.

The “Moral Machine experiment” (Awad et al., 2018) raised the data-driven ethics of AVs to another level. In a multilingual study, the authors collected almost 40 million judgments from people in 233 countries in various moral dilemmas of AVs. Based on the *switch dilemma*, the participants were shown unavoidable accident scenarios with two possible outcomes. The actual outcome of the scenario depended on whether the AV continued on the previous course or took evasive action. The dilemmas were presented to the participants graphically, but a written description of each scenario was available as well. The participants were able to select their preferred maneuver of the AV directly in the graphical representation. The study was conducted online via a dedicated website (the “Moral Machine”) and participants volunteered to take part in the study. Each participant could take part in the study as many times as he or she wished.

In total, the study was based on almost 26 million distinct accident scenarios in which nine factors of the dilemma were systematically varied (e.g., passengers of the AV vs. pedestrians, more victims vs. fewer victims, men vs. women, young vs. old people, humans vs. animals, etc.). Each participant was randomly assigned to 13 scenarios for assessment. In all scenarios, the respective characteristics of the potential victims and the corresponding consequences of the accidents were known with certainty. Uncertainties were not included in the scenarios. In addition to their assessment of the scenarios, participants were able to provide some demographic data in a post-experimental questionnaire and they were geolocated.

From a global perspective, three moral preferences turned out to be strong in the study. These were favoring people over animals, more over fewer people, and younger over older people when deciding which casualties should be spared by the AV. Demographic characteristics of the participants had hardly any noticeable influence in this context. Using hierarchical cluster analysis, three moral clusters emerged at the country level: a Western cluster consisting mainly of countries in North America and Europe, an Eastern cluster with Far Eastern, Confucian and Islamic countries and a Southern cluster with countries mainly from Central and South America and the French overseas territories. The main difference between these clusters was the strength of some preferences. The preference for sparing younger people over older people, for example, was less pronounced in the Eastern cluster and more pronounced in the Southern cluster. In the Southern cluster, the preference for sparing humans over animals was not very strong. According to Awad et al. (2018), the results of the cluster analysis demonstrate that it may be possible to converge on shared preferences for the ethics of AVs within clusters, while this may be more difficult between clusters.

3 Critique of the empirical ethics of AVs

3.1 In the philosophical literature

In the philosophical literature, the empirical ethics of AVs is criticized in many different ways. Some of the criticism disputes that moral intuitions of the general public should matter at all in the development of AVs (see, e.g., Nyholm, 2018; Harris, 2020; Lundgren, 2020; Kochupillai, Lütge and Poszler, 2020). In the broadest sense, this critique ultimately calls experimental ethics itself into question. “Folk intuitions” in artificial thought experiments are hardly transferable to real-life situations (Kochupillai, Lütge and Poszler, 2020) and questions about life and death should therefore generally not be based on gut feelings of people in vignettes (Harris, 2020). Moreover, empirical studies on moral intuitions only collect preferences between different choice options, whereas in an ethical discussion it is essential to formulate and evaluate arguments for or against each option (see, e.g., Nyholm, 2018). In addition, when surveying preferences, there is a risk that these are ethically bad – for example, racist – and thus disqualify themselves as a basis for moral conduct of AVs (Lundgren, 2020). Finally, the general criticism regarding the assessment of moral intuitions sometimes refers to the fact that most participants have little experience with AVs at the time of the study. However, once people gain experience with AVs, their attitudes towards AVs’ driving behavior are likely to change and therefore current attitudes are not very meaningful and only of little use (see, e.g., Nyholm, 2018; Kochupillai, Lütge and Poszler, 2020).

Another strand of criticism questions whether unavoidable accidents in road traffic really entail trolley-like problems. Davnall (2020), for example, argues that the focus on trolley problems in the empirical ethics of AVs is misguided because the best course of action in unavoidable accident situations would always be the initiation of emergency braking. Weighing up different options and possible evasive action, as suggested in trolley problems, would never be a recommended alternative due to driving dynamics. Therefore, moral intuitions in trolley problems are irrelevant for AVs. Himmelreich (2018)

raises a similar argument. If the AV can still weigh options against each other and start a controlled evasive maneuver, this can hardly be an unavoidable accident situation. If an accident is indeed unavoidable, it is unlikely that the AV will be able to take controlled evasive action. The simultaneous occurrence of these two conditions (i.e., unavoidability and control) as a prerequisite for the appearance of real trolley problems in road traffic is therefore rather unlikely. Moreover, according to Himmelreich (2018), trolley problems address the wrong level. Questions about the desired functioning of AVs are ultimately located at the political level. It is about seeking solutions with broad societal acceptance. Questions about preferred actions in trolley problems, on the other hand, are located at the level of morality. Here it is unlikely to find a broad consensus in society, as the right action in trolley problems is highly controversial.

Yet another strand of criticism relates to the discrepancy between trolley problems and real accident situations in road traffic. Trolley problems are based on simple, idealized scenarios that are not representative of the complexity of reality (e.g., Nyholm and Smids, 2016; Lundgren, 2020; Kochupillai, Lütge and Poszler, 2020). In particular, trolley problems, in contrast to real accident situations, are based only on binary choices (e.g., Nyholm and Smids, 2016; Nyholm, 2018; Lundgren, 2020; Kochupillai, Lütge and Poszler, 2020) and they ignore risks and uncertainties (e.g., Nyholm and Smids, 2016; Himmelreich, 2018; Nyholm, 2018; Lundgren, 2020; Kochupillai, Lütge and Poszler, 2020). The latter point is emphasized in almost all critical discussions on the empirical ethics of AVs. Trolley problems are based on fully deterministic scenarios where the outcomes of the choices are known with certainty. Of course, this is not the case in real accident situations. The conditions of the road and weather, the size, weight and speed of other vehicles involved in the accident, the state of health of other drivers or pedestrians, etc., all have an influence on the outcome of the accident and none of this will be known with certainty to AVs. The problem for the empirical ethics of AVs here is that moral preferences in idealized scenarios do not necessarily correspond to those in

real accident situations (see, e.g., Lundgren, 2020). Especially the omission of risks and uncertainties in trolley problems has been suspected of fundamentally changing moral judgments (e.g., Nyholm and Smids, 2016; Himmelreich, 2018; Nyholm, 2018; Lundgren, 2020; Kochupillai, Lütge and Poszler, 2020). Moral judgments in situations of risk and uncertainty are presumably categorically different from moral judgments in deterministic scenarios where everything is known with certainty (Nyholm and Smids, 2016).

The validity and relevance of the respective criticisms are extensively debated in the philosophical literature (see, e.g., Lütge, Rusch and Uhl, 2014; Lin, 2016; Keeling, 2020; Wolkenstein, 2018; Nyholm, 2023). This doctoral thesis addresses the latter criticism and examines to what extent moral judgments change when risks are taken into account in trolley problems. Chapter 4 presents the results of the studies conducted in this regard and Chapter 5 puts the results into context and discusses them in relation to the above-mentioned criticism.

3.2 In the engineering literature

In the engineering literature, the trolley problem is widely rejected. The prevailing consensus is that the trolley problem "should not influence designs for driverless cars or any other autonomous systems" (Winfield et al., 2019). On the one hand, it is pointed out that with the current state of the art, it would simply be impossible to develop an AV that could reliably choose between two unethical outcomes in the dynamic environment of real road traffic (Winfield et al., 2019). On the other hand, trolley problems are considered irrelevant among engineers because they are either implausible or occur only very rarely, if at all (see, e.g., Goodall, 2019). They are considered implausible primarily because trolley problems are based on deterministic scenarios in which all information is known with certainty. However, information from sensors inevitably contains noise (Trussell, 2018). Since trolley problems are considered to be unlikely edge cases, a focus

on them is often even regarded as a waste of resources (Goodall, 2019). Trolley problems in real road traffic can be solved best by trying to avoid them (Johansson and Nilsson, 2016).

From an engineering perspective, “the ultimate desired performance outcome [of AVs] is the ability to drive safely and smoothly through traffic” (Thornton et al., 2016). Safety here basically refers to the avoidance of collisions. Collision avoidance is typically regarded as a “deontological rule” (Thornton et al., 2016) because it is “arguably the highest priority of the automated vehicle” (p. 1432). This is also reflected in the fact that in the engineering literature the AV to be steered is the “ego-vehicle,” while all other vehicles are simply referred to as “obstacles” (Claussmann et al., 2020). The engineering literature on AVs is then concerned with making sure that the “ego-vehicle” avoids collisions with obstacles in road traffic (see, e.g., Reichardt and Shick, 1994; Gehrig and Stein, 2007; Erlien, Fujita and Gerdes, 2013; Wolf and Burdick, 2008; Keller et al., 2014; Funke et al. 2015; Wachenfeld et al., 2016; Gerdes and Thornton, 2016; Thornton et al., 2016; Claussmann et al., 2020). Considerations such as those in the trolley problem do not play a role here.

In a nutshell, the empirical ethics of AVs has so far been concerned with the distribution of harm in accident situations, whereas the engineering science of AVs is about collision avoidance. These are two completely different objectives. From an engineering perspective, it makes more sense to focus on collision avoidance than on the distribution of uncertain outcomes in the event of a collision (Gerdes and Thornton, 2016).

4 Conducted research projects

4.1 Autonomous vehicles and moral judgments under risk[1]

4.1.1 Motivation and aim of the research project

As described in Chapter 2, the empirical ethics of AVs usually revolves around variants of the standard trolley problem (Foot, 1967; Greene, 2013). The relevance and validity of the empirical trolley literature for AVs is, however, often criticized and questioned (see Chapter 3). In this research project, we address two of these common criticisms.

The first criticism relates to the fact that standard trolley problems ignore risks and uncertainties and are therefore unrealistic with respect to road traffic (Goodall, 2016b; Trussell, 2018; Winfield et al., 2019). Each alternative in a typical trolley problem leads to a specific outcome with certainty, but in road traffic outcomes are probabilistic. The problem is, according to the critique, that moral judgments under certainty and under risk (or uncertainty) are categorically different and therefore not necessarily transferable (see, e.g., Fried, 2012; Nyholm and Smids, 2016). The second point of criticism relates to the focus on crash scenarios in trolley problems. The main objective of AVs is to prevent these crash scenarios and not to solve associated ethical dilemmas. A much more important task of AVs therefore is a carefully planned risk management and the associated marginal shifts of risk between different road users (Goodall, 2016a). This distribution of small risks raises different ethical issues than those raised by the trolley problem, according to the critique.

In this research project, we address both points of criticism with the help of online studies. In these studies, we investigate whether people's moral judgments change considerably when they consider (*i*) situations under risk instead of under certainty and (*ii*) situations

[1] This chapter summarizes the first paper of the publication-based dissertation. The research presented here was conducted in collaboration with Matthias Uhl. Details of the publication and the authors' individual contributions can be found in *Appendix: Publication 1*.

with only minuscule accident probabilities instead of situations based on unavoidable crash scenarios.

4.1.2 Methods

In this research project, we conducted three studies with a total of 1,011 participants. Each participant took part in only one of the three studies and was exposed to only one moral dilemma. All participants were residents of the United States and were recruited through CloudResearch Prime Panels (Litman, Robinson and Abberbock, 2017; Chandler et al., 2019). CloudResearch is a platform for sourcing participants for online research. We chose Prime Panels from CloudResearch because they have been shown to provide reliable survey results for a wide range of tasks (Chandler et al., 2019).

Study 1 consisted of two different dilemmas. One was the standard trolley problem, in which a driverless trolley is heading towards five people whose deaths can only be avoided if the trolley is diverted, thereby killing one other person. The other dilemma was a slightly modified version of the standard trolley problem, in which the group of five now had a 1% chance of evading the trolley in time. The group of five thus had a chance of survival, even if it was very small. In Study 2, we gradually increased the chances of survival for the group of five. The initial scenario was again the standard trolley problem. In three further scenarios, we increased the chance of survival for the group of five to 20%, 50% and finally 80% when the trolley approached them.

In Study 3, we further modified the dilemmas to get closer to the actual problems faced by AVs. In these dilemmas, a self-driving car without passengers approached a lane narrowing and could either put a group of five road workers on one side or a single road worker on the other side in danger. The accident probability with the group of five people on one side was either 1.0%, 0.8%, 0.5% or 0.2%, depending on the scenario. The probability of an accident involving the single person on the other side was always 1.0%.

4.1.3 Results

In the standard trolley problem in Study 1, most participants (92%) thought that the single person should be sacrificed for the benefit of the group of five. When the group of five was given a 1% chance of survival, most participants (91%) still thought that the single person should be sacrificed. The two situations under certainty and under risk were apparently assessed identical by the participants. If the chance of survival for the group of five was gradually increased in Study 2, the proportion of participants who would sacrifice the single person gradually decreased. However, despite the vastly reduced accident probability, in none of the three scenarios was there a clear majority who were willing to let the trolley drive towards the group of five. Even with an accident probability of only 0.2 for the group of five, around half of the participants (49%) voted in favor of diverting the trolley to the detriment of the single person.

In Study 3, we adapted the scenarios more closely to road traffic situations and, in addition, significantly reduced the overall accident probability. However, the moral judgments of the participants were virtually identical to those in Study 2. This means that in the case where the probability of an accident was the same for the group of five and the single person, most participants (90%) thought that the single person should bear the risk of an accident. As the accident probability for the group of five decreased, the proportion of participants who would put the single person at risk gradually decreased, just like in Study 2. The participants' moral judgments did not differ between situations with low accident probabilities and those based on unavoidable crash scenarios. The underlying scenario in trolley problems seems to be interchangeable without having much impact on people's moral preferences. Furthermore, the results show that the inclusion of risk in trolley problems does not abruptly change moral judgments. The influence of risk on moral judgments seems to be steady and gradual rather than erratic, as assumed by some critics of the trolley problem.

4.2 Automated vehicles and the morality of post-collision behavior[2]

4.2.1 Motivation and aim of the research project

Imagine a large truck on an interstate highway having an accident with another vehicle and simply driving on. The level of dismay at the truck driver's behavior would be enormous. In many countries, a hit-and-run and the driver's failure to behave appropriately after a collision would even be a criminal offense. Now imagine that the truck was driverless, i.e., a self-driving truck. Would your judgment of the post-collision behavior change? What would you expect how the self-driving truck should behave after a collision?

While appropriate post-collision behavior of a human driver is clearly defined in Article 31 of the 1968 Convention on Road Traffic (UNECE, 2021), there are so far no regulations that define post-collision behavior of AVs. On the contrary, a recent proposed amendment to the Convention on Road Traffic in 2020 proposes to exempt AVs from many articles and clauses related to drivers' behavior requirements, including Article 31 (Economic Commission for Europe, 2020). So far, eCall and Event Data Recorders in AVs are used for occupant protection and will often not even be triggered in the event of collisions with vulnerable road users. An accident involving a pedestrian and a self-driving cab from *Cruise* in October 2023 in San Francisco, in which the cab continued to drive to the side of the road after the collision and pulled the pedestrian along with it, shows the relevance of appropriate post-collision behavior of AVs.

Despite numerous studies on the preferences of AVs for pre-collision behavior, little is known about people's preferences for the capabilities of AVs after a collision. Open questions are, for example, what level of importance people attach to this issue and

[2] This chapter summarizes the second paper of the publication-based dissertation. The research presented here was conducted in collaboration with Matthias Uhl and Bryn Balcombe. Details of the publication and the authors' individual contributions can be found in *Appendix: Publication 2*.

whether car owners would be willing to bear the costs of the necessary technology even if there were no regulatory requirements. Questions such as these are addressed in this study.

4.2.2 Methods

For this study, we conducted an online survey with a total of 1,138 participants from the United States. Participants were recruited via CloudResearch Prime Panels (Chandler, Rosenzweig, Moss, Robinson and Litman, 2019; Litman, Robinson and Abberbock, 2017). As in the previous study, we chose Prime Panels from CloudResearch because they have been shown to provide reliable survey results for a wide range of tasks (Chandler et al., 2019).

In the survey, all participants faced a trolley-like scenario in which an AV had to choose between two courses of action. We used a stochastic trolley problem as the baseline scenario because it explicitly captures important elements of risk management in the context of AVs (see Chapter 4.1). In the baseline scenario, post-collision behavior of the AV did not play a role. In addition to this scenario, we generated two further scenarios in which one of the two options of the AV was coupled with appropriate behavior in the event of a collision and the other was not. In line with Article 31 of the Convention on Road Traffic (UNECE, 2021), appropriate post-collision behavior meant that after an accident, the vehicle (*i*) stops at the site of the accident, (*ii*) calls the police, and (*iii*) records the accident to later determine responsibilities. It should be emphasized that the behavior after the collision had no influence on the outcome of the accident itself. Accidents were fatal regardless of how the AV behaved after the collision.

Each participant in the study was confronted with one of the three scenarios and had to answer two questions about the scenario. First, which of the two possible courses of action the AV should take, and second, an assessment of the relative morality of the two

possible courses of action on a scale ranging from zero to 100. In addition, we elicited participants' normative and empirical expectations regarding the behavior of AVs after a collision in general, and we assessed participants' willingness to pay for the required technology in AVs.

4.2.3 Results

Appropriate post-collision behavior significantly and substantially affected both the choice of the course of action that the AV should choose from the participants' point of view and the assessment of the relative morality of the two possible courses of action. Evidently, appropriate post-collision behavior carried a moral value for the participants. This, in turn, generated spillover effects on the moral evaluation of the underlying crash scenario. The idea that AVs could commit "hit-and-runs" seemed to be daunting for the participants.

Participants accordingly expressed a strong preference that AVs should be able to behave appropriately after a collision and most participants also believed that AVs would have the technological capabilities to do so. The normative and empirical expectations of the participants were closely aligned. Clearly, most participants were unaware that AVs do not currently have these capabilities and may not be required to have them. Participants also expressed a pronounced willingness to pay for the necessary technology in AVs. Overall, therefore, the results clearly indicate that people care a lot about AVs' post-collision behavior and that they place a high value on it.

4.3 The risk ethics of autonomous vehicles: an empirical approach[3]

4.3.1 Motivation and aim of the research project

As already mentioned, road traffic is not deterministic, but is associated with various risks. The main task of AVs is therefore to evaluate and weigh up risks and select one of many possible driving maneuvers. The selected driving maneuver, in turn, has an impact on other nearby road users and distributes traffic risks among them. These risks essentially consist of a specific, possibly very small probability of an accident occurring– for example due to driving errors or technical malfunctions. Standard trolley problems do not capture such considerations.

In this research project, we extend the approach of Chapter 4.1 and explicitly address the possible trade-off between probability and severity of an accident. To this end, we elicited people's moral intuitions in situations involving risk in everyday road traffic. We asked people to adjust safety distances between different road users and thus examined their trade-offs between collision probability and collision severity in road traffic.

4.3.2 Methods

For the study, we developed a slider task. We prepared this task graphically in such a way that it depicts possible traffic situations in a future with AVs (see example in *Figure 1*). In each situation, a self-driving car was depicted between two other road users. The self-driving car represented the slider that could be moved back and forth between the other road users in 99 increments. In this way, the participants were able to set the driving position of the self-driving car and thus the desired safety distances between the other road users. The smaller the distance to another road user, the greater

[3] This chapter summarizes the third paper of the publication-based dissertation. The research presented here was conducted in collaboration with Matthias Uhl. Details of the publication and the authors' individual contributions can be found in *Appendix: Publication 3*.

the probability of a collision with this road user. The overall probability of an accident was lowest when the self-driving car traveled exactly in the middle between the two other road users.

In total, the study consisted of 29 different traffic situations. The situations differed (*i*) in the number of passengers in the cars on the left and right side of the road, (*ii*) whether there was another car or a cyclist on the right side of the road and (*iii*) whether the participants of the study were themselves part of the traffic situation by being passengers in the self-driving car or not. We conducted the study with an online representative sample of 1,807 participants in Germany.

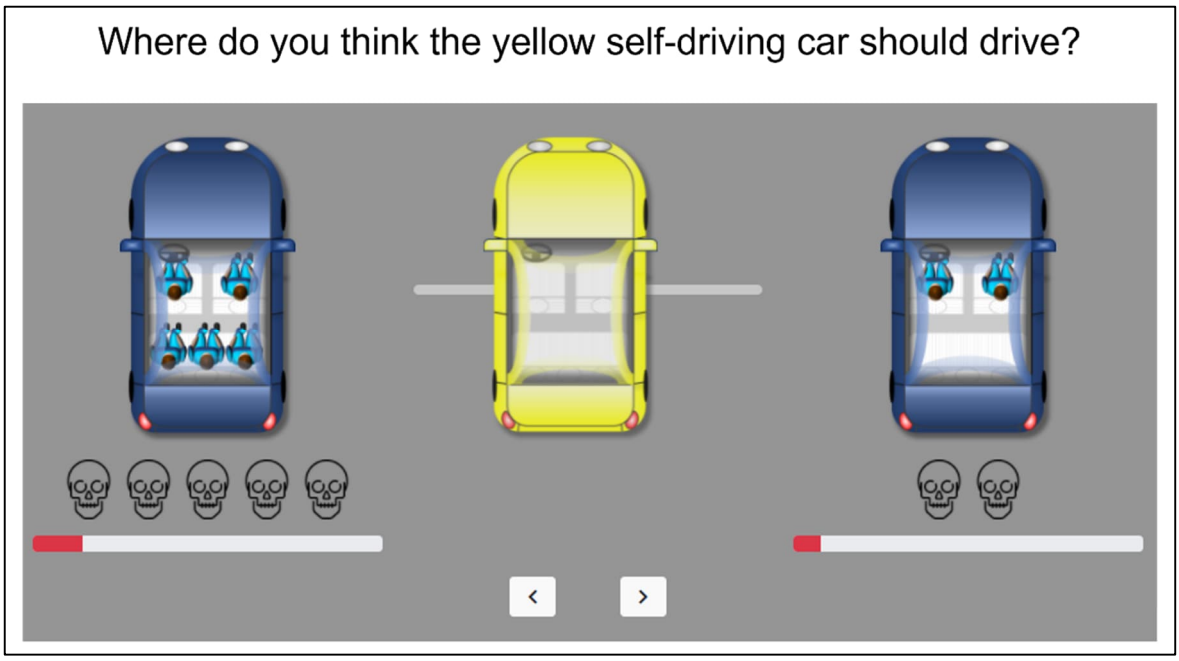

Figure 1: Exemplary illustration of a traffic situation used in the study.
The yellow car in the middle represents the self-driving car, which could be moved back and forth between the other two road users in 99 increments. The smaller the distance to another road user, the greater the probability of a collision with this road user. The red bars below each road user visualized the collision probability with the respective vehicle.

4.3.3 Results

When positioning the self-driving car, the participants paid attention to the number of passengers in the other two vehicles. On average, the self-driving car was always positioned closer to the vehicle with fewer passengers. The more passengers in one of the two vehicles compared to the other, the greater the distance the participants maintained to this vehicle. The participants therefore were not just taking into account the probability of an accident, but also the severity of the accident. In our study, it did not matter whether the participants themselves were part of the traffic situation or not. They were willing to bear traffic risks themselves if it reduced the probability of a more severe accident. This is surprising, as previous studies with deterministic trolley problems have found that people prefer AVs that protect them as passengers at all costs (Bonnefon et al., 2016). When a cyclist was depicted on the right side of the road, the cyclist was given slightly more safety distance than when a vehicle with one passenger was depicted on the right side of the road. This indicates that the participants in our study attributed a small risk bonus to the cyclist, even though this bonus was not very pronounced.

The results demonstrate that it is not only the probability of an accident that might matter to people in road traffic. Our study participants expect that AVs also take into account the severity of an accident in their risk management, for example in terms of the number of possible casualties. This remains true even if the participants themselves were passengers in the AV. Using a follow-up study, we tested and successfully showed that our study participants did indeed understand that the middle driving position of the self-driving car would have minimized the overall probability of an accident.

5 Discussion

5.1 Road traffic as a matter of risk management

Any participation in road traffic entails a risk of accidents for other road users. Accidents can occur due to technical failure, human error or unforeseeable events. The probabilities of an accident may be small, but they are strictly positive. The distribution of these accident probabilities to other road users obviously has an ethical component. An example would be the lateral distance to a bicycle lane that is not structurally separated from car traffic (Bonnefon, Shariff and Rahwan, 2019). Vehicles can position themselves laterally anywhere in a lane as long as they stay within the lane markings (Goodall, 2019). The closer the AV drives to the bicycle lane, the higher the probability of an accident with a cyclist. However, the greater the distance to the bicycle lane, the higher the probability of the AV colliding with oncoming traffic on the other side of the road. Safety distances between different road users are a prototypical example of such risk considerations. A key characteristic is that choices in road traffic must be made under constraints (i.e., limitations), for example the structural limits of the road. It is often just not possible to minimize the probability of an accident with one road user without affecting the probability of an accident with another road user. Other examples include safety distances to leading vehicles and various braking strategies to weigh up the risks of a collision with a leading and a following vehicle (Goodall, 2019).

The empirical ethics of AVs has so far focused on the distribution of harm in unavoidable accident situations. As Goodall (2016a) points out, very similar questions arise in everyday traffic situations such as those described above, in which accident probabilities are shifted between different road users. The question here is not how an AV should choose in the event of an unavoidable accident, but rather the much more fundamental question of how the AV should behave on the road in general to minimize the probability of an accident with a particular road user, while potentially increasing the probability of

an accident with another road user. Questions such as these are of utmost importance, as almost every active intervention in road traffic leads to a redistribution of risks between road users.

Contrary to humans, who make these driving decisions instinctively, AVs would have to do so on the basis of a carefully planned risk management (Goodall, 2016a, b). Whereas the scholarly engineering sciences are mainly concerned with collision avoidance strategies of AVs, some car companies seem to be going further in this context. *Google Inc.* and *Waymo LLC*, for instance, describe in a series of patents how AVs can compare different driving maneuvers using a risk-cost framework (Teller and Lombrozo, 2014, 2015, 2019). The AV calculates the risk of these maneuvers and selects the best maneuver based on an expected-utility criterion. In these patents, other road users are not just “obstacles” to be avoided, but their own risks are explicitly included in the calculations. Teller and Lombrozo (2014, 2015, 2019) also describe the possibility of differentiating according to the type of road user in order to grant pedestrians, for instance, greater consideration in this risk assessment. In another patent, *Google Inc.* explains how an AV can adapt the lateral driving position within a lane to different traffic situations (Dolgov and Urmson, 2014). According to Dolgov and Urmson (2014), an AV could, for instance, increase its own safety by increasing the lateral distance to a truck, even if this reduces the lateral distance to a small car on the other side of the road. However, the patent also describes ways in which the AV could take into account the vulnerability of other road users in this assessment and, for instance, maintain a greater lateral distance to cyclists.

This form of risk management is an important task of AVs (Goodall, 2016a), which must be discussed from an ethical perspective. Should certain road users be granted special treatment, for instance by maintaining greater safety distances? Can occupants of AVs be expected to accept minimal increases in accident probabilities in order to reduce risks for other road users?

5.2 The difficulties of risk ethics

The previous chapter shows that similar issues arise in the risk management of AVs as in unavoidable accident situations. The difference is that many of the criticisms targeting the empirical studies of AVs mentioned in Chapter 3 do not apply to problems of risk management. Risk management is necessary in everyday situations in road traffic. These situations are unavoidable and at the same time controllable. Emergency braking does not offer an easy way out of the dilemma. Furthermore, these situations do not represent edge cases and the patents from *Google Inc.* and *Waymo LLC* clearly show that the car industry does not dismiss potential use cases as implausible. Therefore, scientific discussions on ethical issues in the context of the risk management of AVs do not appear to be a waste of resources.

However, taking risk into account shifts the focus of the discussions. Stochastic trolley problems require different answers than deterministic trolley problems. Philosophy has so far largely ignored the problems that arise from taking risk into account (Hansson, 2012, 2013). One of the most important problems in ethics is extending standard ethical theories to ensure an ethical account of situations under risk (Hansson, 2012). Currently, not even the term “risk” itself has been clarified (see, Hansson, 2012, 2013). According to Hansson (2012, 2013), sometimes risk refers to an unwanted event that may or may not occur (e.g., lung cancer as a risk of smoking), sometimes to the cause of an unwanted event (e.g., smoking as a health risk).

There are different approaches to comparing risks as well. Sometimes risk refers exclusively to the probability of an unwanted event occurring, sometimes to the expected value of the severity of an unwanted event (Hansson, 2012, 2013). The engineering literature, for example, uses the term risk in its first meaning. The unwanted event is a collision with another road user (i.e., the “obstacle”) and risk refers to the probability that the unwanted event will occur. Decision theory usually understands risk in the second meaning, i.e., the probability-weighted severity of the unwanted event. But even this view

is not uncontroversial. According to Hansson (2012), risks with the same expected value are not always considered equally serious. Sometimes it might be preferable to give serious events with a lower probability more weight in the calculation of the expected value.

But even if one agrees on a concept of risk, moral theories quickly reach their limits when trying to account for risk (however defined) (Hansson, 2012, 2013, 2023). For example, if the duty in a deontological theory not to harm other people were to be extended to the duty not to take actions that increase the risk of harming others, social interaction in its current form would hardly be possible (Hansson, 2023). If, on the other hand, risk assessments are based on the expected-utility criterion in a utilitarian theory, a disproportionate weighting of major disasters would not be permissible, although many people would prefer to do so (Hansson, 2023).

The amount of risk that may be imposed on one person in order to protect another person from being exposed to risk involves difficult moral questions, not purely decision-theoretical considerations (Hansson, 2012). It would be a technocratic fallacy to conclude from the fact that engineers and vehicle technicians are able to determine the risk posed by AVs that they are therefore also competent to decide whether this risk is acceptable (see Hansson, 2004). The latter is a question of value judgments that cannot be derived from vehicle technology. Unfortunately, current moral theories do not yet offer a satisfactory solution for the assessment of risk either (Hansson, 2023).

6 Conclusion

The empirical ethics of AVs so far has been based mainly on deterministic trolley problems. The lack of realism of these dilemmas often triggers strong reservations and the relevance of the study results for AVs has often been questioned. In this dissertation, I addressed this criticism and examined the extent to which moral judgments change when the underlying dilemmas are based on decisions under risk. In this regard, it is found that moral judgments do not change categorically just because risks are taken into account. This is particularly then the case when the participants were not part of the underlying dilemma. However, when the participants were part of the dilemma, there was a crucial and surprising difference between stochastic and deterministic trolley problems. In the former, the participants were willing to take risks for the benefit of others. In the latter, this type of altruism was not evident. For the possible acceptance of AVs with utilitarian tendencies, this may make a huge difference.

Overall, our participants expected a balanced consideration between accident probability and accident severity. This runs counter to the current belief of many engineers that the sole focus on accident probability in AV design is not only sensible but also acceptable. However, as Hansson (2004) points out, questions of the acceptability of risk are value judgments that are not easy to answer. Unfortunately, risk ethics does not provide much guidance here at the moment. An obvious and solution-oriented approach would be to ask those people about the acceptability of risks who are themselves affected by them: the general public. This thesis provides a first step in this direction. Methods have been developed and applied with which preferences for the distribution of risks in road traffic can be surveyed. These methods should be developed further and applied in an international context. In this way, it is possible to study how people weight the severity and probability of accidents in road traffic.

Such considerations have not only played a role since the advent of AVs. In Vision Zero concepts, which many countries are now pursuing in road traffic, intersections are

systematically replaced by traffic circles (Belin, Tillgren and Vedung, 2012). The aim of these Vision Zero concepts is reducing the number of traffic fatalities and serious injuries, ideally even avoiding them altogether. These concepts take advantage of the fact that the probability of serious accidents is lower at traffic circles than at intersections due to lower speeds (Belin, Tillgren and Vedung, 2012). Proponents of Vision Zero concepts believe that this is justified, even though the probability of accidents at traffic circles is higher overall. Irrespective of how one may think about this approach, the example makes it clear that the risk to be distributed in road traffic does not have to be managed by AVs alone. It is a matter of the design of road traffic as a whole.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729), 59-64.

Belin, M.Å., Tillgren, P., Vedung, E. 2012. Vision Zero–a road safety policy innovation. *International journal of injury control and safety promotion*, 19(2), 171-179.

Bigman, Y.E., Gray, K. 2020. Life and death decisions of autonomous vehicles. *Nature*, 579(7797), E1-E2.

Bonnefon, J.F., Shariff, A., Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573-1576.

Bonnefon, J.F., Shariff, A., Rahwan, I. 2019. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, 107(3), 502-504.

Bote, J. 2023. *Waymo driverless car brings San Francisco traffic to a halt during rush hour*. SFGate, January 24. https://www.sfgate.com/bayarea/article/waymo-rush-hour-traffic-standstill-17739556.php

Cano, R. 2024. *One crash set off a new era for self-driving cars in S.F. Here's a complete look at what happened*. San Francisco Chronicle, February 8. https://www.sfchronicle.com/projects/2024/cruise-sf-collision-timeline/

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., Litman, L. 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5), 2022-2038.

Claussmann, L., Revilloud, M., Gruyer, D., Glaser, S. 2019. A review of motion planning for highway autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(5), 1826-1848.

Connatser, M. 2024. *Waymo robotaxi drives down wrong side of street after being alarmed by unicyclists: Strange tales from San Francisco*. The Register, April 23. https://www.theregister.com/2024/04/23/waymo_selfdriving_car_unicycle/

Davnall, R. 2020. Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics. *Science and Engineering Ethics*, 26(1), 431-449.

Dolgov, D., Urmson, C. 2014. *Controlling vehicle lateral lane positioning* (U.S. Patent No. 8,781,670 B2). U.S. Patent and Trademark Office.

Economic Commission for Europe. 2020. *Revised Amendment proposal to the 1968 Convention on Road Traffic*. Economic Commission for Europe, Inland Transport Committee, Global Forum for Road Traffic Safety. https://unece.org/fileadmin/DAM/trans/doc/2020/wp1/ECE-TRANS-WP1-2020-1-_Rev1e_.pdf

Erlien, S. M., Fujita, S., Gerdes, J. C. 2013. Safe driving envelopes for shared control of ground vehicles. *IFAC Proceedings Volumes*, 46(21), 831-836.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., König, P. 2019. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and engineering ethics*, 25, 399-418.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 1-7.

Frank, D. A., Chrysochou, P., Mitkidis, P., Ariely, D. 2019. Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific reports*, 9(1), 13080.

Fried, B.H., 2012. What does matter? The case for killing the trolley problem (or letting it die). *The Philosophical Quarterly*, 62(248), 505-529.

Funke, J., Brown, M., Erlien, S. M., Gerdes, J. C. 2015. Prioritizing collision avoidance and vehicle stabilization for autonomous vehicles. In: *2015 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 1134-1139.

Gehrig, S. K., Stein, F. J. 2007. Collision avoidance for vehicle-following systems. *IEEE transactions on intelligent transportation systems*, 8(2), 233-244.

Gerdes, J. C., Thornton, S. M. 2016. Implementable Ethics for Autonomous Vehicles. In: Maurer, M., Gerdes, J., Lenz, B., Winner, H. (eds.), *Autonomous driving: Technical, legal and social aspects*, Springer: Berlin, Heidelberg, 87-102.

Goodall, N.J., 2016a. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6), 28-58.

Goodall, N.J., 2016b. Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, 30(8), 810-821.

Goodall, N. 2019. More than trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles. *Transfers*, 9(2), 45-58.

Greene, J.D. 2013. *Moral tribes: Emotion, reason, and the Gap between us and them*. Atlantic Books: London.

Greene, J. D. 2016. Our driverless dilemma. *Science*, 352(6293), 1514-1515.

Hansson, S.O. 2004. Fallacies of risk. *Journal of Risk Research*, 7(3), 353-360.

Hansson, S.O. 2012. A panorama of the philosophy of risk. In: Roeser, S., Hillerbrand, R., Sandin, P., Peterson, M. (eds), *Handbook of risk theory*, Springer: Dordrecht, Heidelberg, 27-54.

Hansson, S.O. 2013. *The ethics of risk: Ethical analysis in an uncertain world*. Palgrave Macmillan: New York.

Hansson, S. O. 2023. Risk. *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Zalta, E. N., Nodelman, U. (eds.), https://plato.stanford.edu/archives/sum2023/entries/risk/.

Harris, J. 2020. The immoral machine. *Cambridge Quarterly of Healthcare Ethics*, 29(1), 71-79.

Himmelreich, J. 2018. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, 21(3), 669-684.

Huang, K., Greene, J. D., Bazerman, M. 2019. Veil-of-ignorance reasoning favors the greater good. *Proceedings of the national academy of sciences*, 116(48), 23989-23995.

Johansson, R., Nilsson, J. 2016. Disarming the trolley problem–why self-driving cars do not need to choose whom to kill. In: *Workshop CARS 2016-Critical Automotive applications: Robustness & Safety*.

Keeling, G. 2020. Why trolley problems matter for the ethics of automated vehicles. *Science and engineering ethics*, 26(1), 293-307.

Keller, M., Hoffmann, F., Hass, C., Bertram, T., Seewald, A. 2014. Planning of optimal collision avoidance trajectories with timed elastic bands. *IFAC Proceedings Volumes*, 47(3), 9822-9827.

Kochupillai, M., Lütge, C., Poszler, F. 2020. Programming away human rights and responsibilities? “The moral machine experiment” and the need for a more “humane” AV future. *NanoEthics*, 14(3), 285-299.

Lin, P. 2016. Why ethics matters for autonomous cars. In: Maurer, M., Gerdes, J., Lenz, B., Winner, H. (eds.), *Autonomous driving: Technical, legal and social aspects*, Springer: Berlin, Heidelberg, 69-85.

Litman, L., Robinson, J., Abberbock, T. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433-442.

Lundgren, B. 2021. Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles. *Ai & Society*, 36(2), 405-415.

Lütge, C. 2017. The German ethics code for automated and connected driving. *Philosophy & Technology*, 30, 547-558.

Lütge, C., Rusch, H., Uhl, M. 2014. *Experimental ethics: Toward an empirical moral philosophy*. Palgrave Macmillan: New York.

Meder, B., Fleischhut, N., Krumnau, N. C., Waldmann, M. R. 2019. How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk analysis*, 39(2), 295-314.

Morita, T., Managi, S. 2020. Autonomous vehicles: Willingness to pay and the social dilemma, *Transportation Research Part C: Emerging Technologies*, 119, 102748.

Nyholm, S. 2018. The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, 13(7), e12507.

Nyholm, S. 2023. Ethical accident algorithms for autonomous vehicles and the trolley problem. In: Lillehammer, H. (ed.). *The Trolley Problem*, Cambridge University Press: Cambridge, 211-230.

Nyholm, S., Smids, J. 2016. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical theory and moral practice*, 19(5), 1275-1289.

Reichardt, D., Shick, J. 1994. Collision avoidance in dynamic environments applied to autonomous vehicle guidance on the motorway. In: *Proceedings of the Intelligent Vehicles' 94 Symposium*, IEEE, 74-78.

Teller, E., Lombrozo, P. 2014. *Consideration of risks in active sensing for an autonomous vehicle* (U.S. Patent No. 8,781,669 B1). U.S. Patent and Trademark Office.

Teller, E., Lombrozo, P. 2015. *Consideration of risks in active sensing for an autonomous vehicle* (U.S. Patent No. 9,176,500 B1). U.S. Patent and Trademark Office.

Teller, E., Lombrozo, P. 2019. *Consideration of risks in active sensing for an autonomous vehicle* (U.S. Patent No. 10,427,672 B2). U.S. Patent and Trademark Office.

Thornton, S. M., Pan, S., Erlien, S. M., Gerdes, J. C. 2016. Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 18(6), 1429-1439.

Trussell, H.J. 2018. Why a special issue on machine ethics. *Proceedings of the IEEE*, 106(10), 1774-1776.

UNECE. 2021. Road traffic and road signs and signals agreements and conventions. https://unece.org/road-traffic-and-road-signs-and-signals-agreements-and-conventions

Wachenfeld, W., Winner, H., Gerdes, J. C., Lenz, B., Maurer, M., Beiker, S., Fraedrich, E., Winkle, T. 2016. Use cases for autonomous driving. In: Maurer, M., Gerdes, J., Lenz, B., Winner, H. (eds.), *Autonomous driving: Technical, legal and social aspects*, Springer: Berlin, Heidelberg, 9-37.

Winfield, A.F., Michael, K., Pitt, J., Evers, V. 2019. Machine ethics: the design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE*, 107(3), 509-517.

Wolf, M. T., Burdick, J. W. 2008. Artificial potential functions for highway driving with collision avoidance. In: *2008 IEEE International Conference on Robotics and Automation*, IEEE, 3731-3736.

Wolkenstein, A. 2018. What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology*, 20(3), 163-173.

Appendix

Publication 1

Krügel, S., Uhl, M. (2022). Autonomous vehicles and moral judgments under risk. *Transportation Research Part A: Policy and Practice* 155, 1–10.

Available at: https://doi.org/10.1016/j.tra.2021.10.016

Author contributions: S.K. and M.U. developed the research question, designed and performed the studies, and wrote the research article together. S.K. analyzed the data.

Publication 2

Krügel, S., Uhl, M., Balcombe, B. (2021). Automated vehicles and the morality of post-collision behavior. *Ethics and Information Technology* 23, 691–701.

Available at: https://doi.org/10.1007/s10676-021-09607-w

Author contributions: S.K., M.U. and B.B. developed the research question. S.K. and M.U. designed and performed the study, and wrote the research article together. S.K. analyzed the data. B.B. reviewed and approved the manuscript.

Publication 3

Krügel, S., Uhl, M. (2024). The risk ethics of autonomous vehicles: an empirical approach. *Scientific Reports* 14, 1–11.

Available at: https://doi.org/10.1038/s41598-024-51313-2

Author contributions: S.K. and M.U. developed the research question, designed and performed the study, and wrote the research article together. S.K. analyzed the data.