Sci Eng Ethics
https://doi.org/10.1007/s11948-018-0020-x

ORIGINAL PAPER

Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles

Anja K. Faulhaber[1] · Anke Dittmer[1] · Felix Blind[1] · Maximilian A. Wächter[1] · Silja Timm[1] · Leon R. Sütfeld[1] · Achim Stephan[1] · Gordon Pipa[1] · Peter König[1,2]

Received: 24 August 2017 / Accepted: 10 January 2018

Abstract Ethical thought experiments such as the trolley dilemma have been investigated extensively in the past, showing that humans act in utilitarian ways, trying to cause as little overall damage as possible. These trolley dilemmas have gained renewed attention over the past few years, especially due to the necessity of implementing moral decisions in autonomous driving vehicles (ADVs). We conducted a set of experiments in which participants experienced modified trolley dilemmas as drivers in virtual reality environments. Participants had to make decisions between driving in one of two lanes where different obstacles came into view. Eventually, the participants had to decide which of the objects they would crash into. Obstacles included a variety of human-like avatars of different ages and group sizes. Furthermore, the influence of sidewalks as potential safe harbors and a condition implicating self-sacrifice were tested. Results showed that participants, in general, decided in a utilitarian manner, sparing the highest number of avatars possible with a limited influence by the other variables. Derived from these findings, which are in line with the utilitarian approach in moral decision making, it will be argued for an obligatory ethics setting implemented in ADVs.

Keywords Autonomous driving · Utilitarianism · Trolley problem · Moral dilemma

Anja K. Faulhaber, Anke Dittmer, Felix Blind, Maximilian A. Wächter and Silja Timm: Shared first authorship.

✉ Maximilian A. Wächter
mwaechter@uni-osnabrueck.de

[1] Institute of Cognitive Science, University of Osnabrück, Wachsbleiche 27, 49090 Osnabrück, Germany

[2] Department of Neurophysiology and Pathophysiology, Center of Experimental Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Published online: 22 January 2018

Introduction

Since their invention in the nineteenth century, cars have considerably influenced townscapes and societies all over the world. Due to the continuous development and increasing sophistication of vehicles, this impact is still ongoing. It even seems that car manufacturers are getting closer to reaching another milestone: a car that is capable of driving without a human driver. In the last few years, there have been substantial advancements in the development of such Autonomous Driving Vehicles (ADVs). Many features of automation, such as cruise control, camera-based blind spot assistance, and parallel parking have already become standard in modern cars. The majority of car manufacturers and service providers, such as Uber, are currently working on ADVs and planning to commercially market them by 2025 at the latest (Hars 2016). The fast introduction of ADVs is due to the expected advantages. These might include higher mobility for people unable to drive a car (e.g. elderly, tired, disabled people), better organized traffic and fewer traffic jams due to communication between vehicles. Most importantly, based on improved driving behavior and shorter reaction times the number of traffic accidents and casualties is expected to decrease significantly. However, with the development of disruptive technologies, new problems arise. Because introducing ADVs might have a large impact on society, critical issues spread over a wide range of areas including psychological, ethical, socioeconomic, and legal aspects.

The most pressing issues that need to be addressed include liability in the case of casualties as well as the ADV's behavior in moral dilemma situations. Moral decision-making seems to have little implication for traffic so far because most accidents happen in a split second without the time and the information to think thoroughly about one's reaction. Therefore, humans base their decisions in such situations mostly on reflexes and instincts rather than deep thoughts. This will change with the introduction of ADVs given that the car's decisions in all kinds of possible traffic scenarios will be programmed beforehand including guidelines for unforeseen events and even highly unlikely scenarios (Lin 2013, 2015). But there is no consensus yet on who decides what should be programmed. One possibility would be that users may choose an individual ethics setting themselves. However, Gogoll and Müller (2017) criticized and rejected this option as this would most likely lead to a prisoner's dilemma[1] in traffic (see also "Discussion" section). In their thought experiment, people would choose a suboptimal and thus a negative outcome for the entire society, just to prevent a possible exploitation by other road users. In consequence, Gogoll and Müller call for an ethics setting that is mandatory for all ADVs. For

[1] The prisoner's dilemma is a mathematical theory based on game theory. Imagine two prisoners accused of committing a crime together. The two prisoners are interrogated and can not communicate with each other. If both deny the crime, both receive a low punishment. If both are confessing both receive a heavy sentence. However, if only one of the two prisoners confesses, he or she leaves the court without a sentence, while the other gets the maximum sentence. The dilemma in this situation is, that every prisoner must choose to either deny or confess without knowing the other prisoner's decision. The sentence depends on how the two prisoners testify together, and thus depends not only on their own decision but also on the decision of the other prisoner.

the implementation of such a setting, an ethical framework is needed which remains widely debated (Hevelke and Nida-Rümelin 2014, 2015a, b). One problem is that people are in favor of ADVs programmed in a utilitarian way but state they would themselves not want to buy such an ADV (Bonnefon et al. 2016).

Moral decisions by autonomous systems are often discussed on the basis of trolley dilemmas. The classical trolley dilemma was introduced in 1967 as a philosophical thought experiment (Foot 1967). The key element is a trolley heading straight toward a group of people, who are on the rails and unable to escape. There is, however, a side track on which a single person stands, unaware of the trolley. Participants in this thought experiment are standing next to a lever that enables the trolley to switch to the side track, resulting in a moral dilemma. Without intervention, the trolley will kill the group of people on the main track. Upon pulling the lever, the trolley will continue on the side track, killing only one person. How do people make decisions in such situations and what moral principles govern their decision process? This question has been investigated and debated extensively (Mikhail 2007; Thomson 1976, 1985; Unger 1996).

So far, research on modified trolley dilemmas in the context of ADVs focuses on whether there is a moral argument for ADVs to act in a deontological or utilitarian way. The distinction between the ethical theory concepts of deontological motivations and utilitarian motivations is hard to draw, especially with a broad notion of deontology. We define utilitarian actions, opposed to random behavior or refusal of behavior, as those maximizing utility by seeking to cause as little overall damage as possible, based on some probabilistic view of the future. This might even include willingness to risk harm for oneself. Recent studies showed that people in general act in utilitarian ways and are relatively comfortable with utilitarian ADVs, programmed to minimize harm (Bonnefon et al. 2016; Skulmowski et al. 2014).

Besides the deontological approach and utilitarianism, there are many more ethical standpoints concerning how to tackle the problem of decision-making in self-driving cars. These range from virtue ethics, meaning that specialists and ethicists influence the decision-making in machines as a governmental committee, to a machine that mimics the entire spectrum of human behavior (Goodall 2014). Within these discussions, there is much disagreement in the literature about which ethical setting is the right one to choose and no clear statement could be made yet (Lin 2015). For the present study, the behavior of the participants served as a starting point. The intention is to deduce rules from human behavior that would be applicable to all ADVs. This is because people have to agree to an ethical setting that is implemented in their car to actually use it. Moreover, people usually do not judge a case based on deontic or utilitarian grounds but are rather guided by normative standards from their culture and society. This study consequently aims to establish an ethical decision-making framework for moral dilemmas in driving situations that can then serve as a foundation for an obligatory ethical setting to be implemented in ADVs. Such a framework should prevent a system, able to save thousands of lives, not being used because of moral disagreements with the general population.

Studies including trolley dilemmas were traditionally carried out in the form of philosophical essays. This means that the material was presented to participants in the form of written scenario descriptions, sometimes with additional pictorial

representations. This way of presenting the dilemma introduces issues, such as the disregard of important contextual and situational influences in moral decision-making (Skulmowski et al. 2014). New immersive technologies, such as Virtual Reality (VR), could help to remedy these insufficiencies. In this context, trolley dilemmas have recently experienced a revival in science (Navarrete et al. 2012; Pan et al. 2011; Patil et al. 2014; Skulmowski et al. 2014). The immersion that VR environments provide serves to improve ecological validity while maintaining control over experimental variables (Madary and Metzinger 2016). In the context of ADVs, VR can present scenarios that are more similar to real life decision-making in traffic and hence shed light on the moral actions of the participants rather than their conscious beliefs.

Furthermore, many possible modifications of the trolley dilemma elicit open questions. For example, different characteristics of potential victims might influence the human decision process. Previous studies have shown that children were saved more often than adults, so the ages of potential victims might play a role in decision-making (Sütfeld et al. 2017). In the context of ADVs, there are also certain traffic-specific aspects worth considering. For instance, sidewalks provide a safe space for pedestrians in traffic which might lead to an internalized reluctance to drive on sidewalks and could also influence the decision process in a modified trolley dilemma. Additionally, there are possible scenarios in which people can only save lives by sacrificing their own. Despite evidence from surveys that revealed a willingness to use self-sacrificing ADVs (Bonnefon et al. 2016), it is questionable whether people would indeed act this way in realistic settings. The present study specifically addresses these open questions and aims at a high ecological validity by using a VR setting.

In this experiment, five hypotheses were tested. First, based on previous research, it is postulated that people will, in general, act in favor of the quantitative greater good, trying to keep the number of persons to be hit to a minimum (Hypothesis 1). Yet, it can be speculated that the ages of potential victims matter in the sense that people might spare younger individuals at the expense of older ones (Hypothesis 2). In the traffic-specific context pedestrians on the sidewalk are expected to be protected, as they are not actively taking part in traffic. By staying on the sidewalk, people generally expect to be safe while implicitly giving consent to the finite risk of being injured when stepping into the street. Therefore, people are assumed to avoid hitting pedestrians on sidewalks as opposed to people standing on streets (Hypothesis 3). On the other hand, it is hypothesized that people prefer to protect children, even if they are standing on streets, as opposed to adults on sidewalks (Hypothesis 4). Finally, the last hypothesis states that people will not reject self-sacrifice completely but consider it when a high threshold of damage to others is reached (Hypothesis 5). To test these hypotheses, a driving simulation experiment with state-of-the-art VR technologies was implemented, following a study by Sütfeld et al. (2017). The presented avatars were only male to avoid an effect of gender difference, as previous studies showed that male and female avatars are treated differently (Sütfeld et al. 2017). Participants were able to control cars as drivers and experienced various modified trolley dilemma situations, as specified in "Materials and Methods" section.

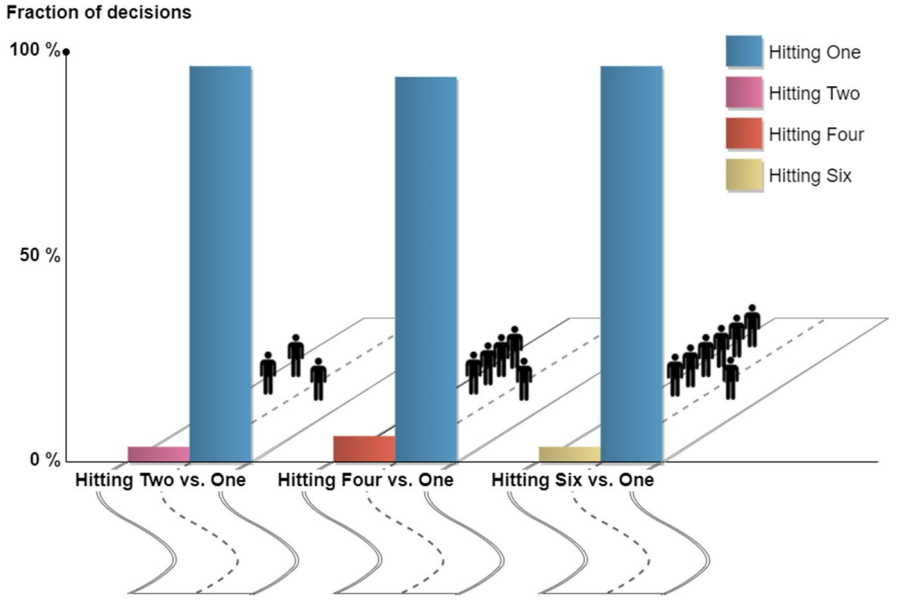

Fig. 1 Decision distribution in the Quantitative Greater Good module

Results

Data from 189 participants and a total of 4000 trials, distributed into five modules according to the aforementioned hypotheses, was analysed. Below, the results for each module will be described separately.

Quantitative Greater Good

In the first module, it was tested whether people would act in favor of the quantitative greater good by saving more as opposed to fewer avatars. This module consisted of three trials. The environment for this module was a suburban setting, consisting of a two-lane road. Only standing adults were presented as avatars. In the suburban setting, parked cars occupied both sides of the two-lane street. In the one-versus-two and one-versus-six conditions, only 7 out of 189 participants targeted the higher number of avatars (Fig. 1). In the one-versus-four condition, 12 participants targeted the group of four instead of the individual; thus, in all three conditions, the overwhelming majority of participants spared the larger number of avatars.

To investigate this difference between the conditions, a permutation test was used. It yielded no significant difference ($p > 0.05$). This shows that participants acted similarly throughout all three conditions. For each single condition, the number of participants targeting one avatar instead of the larger number is

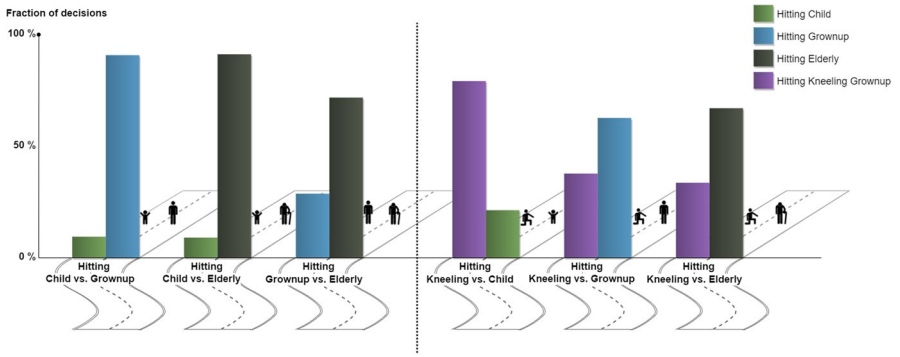

Fig. 2 Decision distribution in the Age-Considering Greater Good module. The left side shows purely age-considering decisions; the right side shows decisions about object height

highly significant ($p < 0.01$). This data indicates that participants decided in favor of the quantitative greater good.

Age-Considering Greater Good

The second module tested the hypothesis that people would spare younger avatars at the expense of older ones. It was composed of six trials in a suburban setting. As avatars a child, a standing adult, a kneeling adult, and an old person were used. Each trial presented one of the following six combinations of avatars: one child versus one standing adult, one child versus one old person, one standing adult versus one old person, one kneeling adult versus one standing adult, one kneeling adult versus one old person, and one kneeling adult versus one child.

In the pairwise comparisons of children, adults, and the elderly, it was observed that younger avatars were spared at the expense of older avatars (Fig. 2). The differences between children versus adults and elderly versus adults were highly significant in a permutation test ($p < 0.001$). The results demonstrate the inverse relation of the expected remaining lifespan of an avatar and the chance of getting hit. This decrease in value according to age was highly significant ($p < 0.01$).

To investigate whether the difference emerged only through variation in avatar height, kneeling adults versus standing children and standing elderly were tested. The observed difference in the children versus kneeling adults' comparison was highly significant (Fig. 2, 4th block, $p < 0.001$). In the direct comparison of kneeling adults versus standing adults, the latter were hit more often ($p < 0.001$). A similar pattern emerged in the comparison between kneeling adults versus the elderly; thus, kneeling and standing moderated the participants' decisions to some degree. However, these results confirm that participants spare younger avatars at the expense of older ones, irrespective of the avatars' heights.

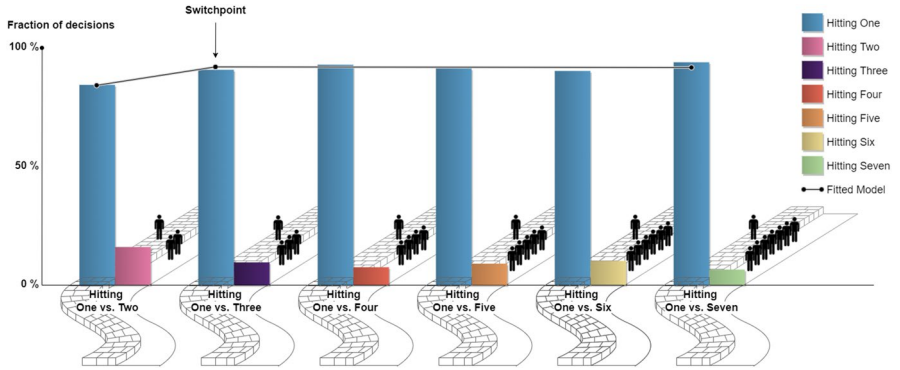

Fig. 3 Decision distribution in the Influence of Context module. Depiction of the best-fitted model for these decisions

The Influence of Context

The third module explored the influence of context. Specifically, the corresponding hypothesis states that avatars located on sidewalks would be spared more often than those in streets. Therefore, in direct analogy to the first module, a single adult avatar on the sidewalk was matched with two to six adult avatars in the street.

This module consisted of six trials in a city setting that contained a one-way street with sidewalks on both sides. One of the sidewalks was blocked by parked cars while participants had the opportunity to drive on the other sidewalk to avoid avatars in the street.

Compared to the first module, it was expected that a larger difference in the number of avatars would be necessary to lead to a consistent sacrifice of the single avatar on the sidewalk. However, in general, this context did not seem to have a strong effect on decisions. The majority of participants still consistently spared the highest number of avatars possible, regardless of the sidewalk context (Fig. 3). It was investigated whether a switch point, defined by a critical imbalance of the number of avatars, could adequately describe the participants' decisions. That is, if the number of avatars to be hit in the street were larger than this threshold, participants would change from driving in the street to driving on the sidewalk to save a large enough group of avatars. The data showed that only 2.56% of trials would need to be changed for all participants to behave consistently according to a simple model with a single free parameter, the switch point.

For statistical evaluation, models describing different switch points were fitted to the data and compared the sums of squared residuals of the models to identify the model that best fits the data. Results showed that modeling the data with a switch point between the conditions with one-versus-two and one-versus-three avatars described the data best (Fig. 3), with a sum of squared residuals of 34.0. This, in turn, indicates that participants choose to drive on the sidewalk to save a group of three or more avatars rather than saving only two. However, throughout all conditions, the number of participants who drove on the sidewalk to save

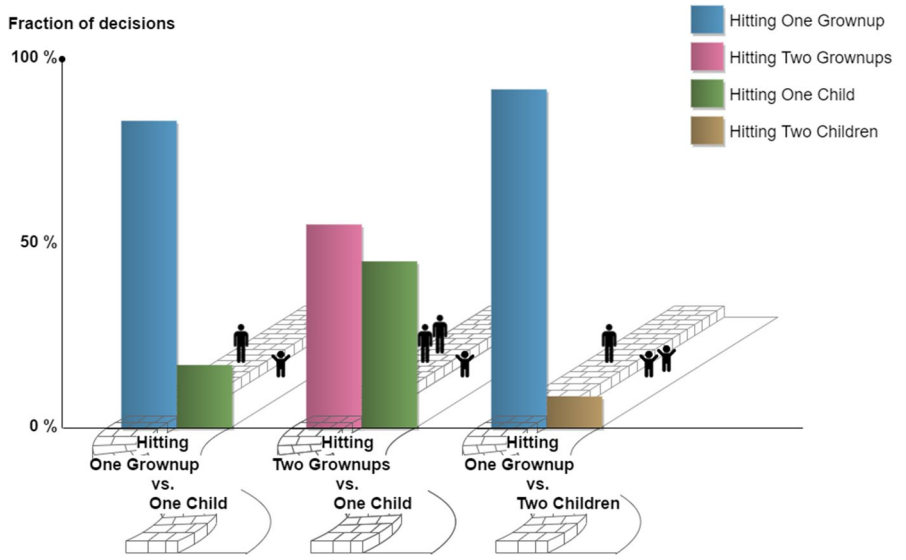

Fig. 4 Decision distribution in the Interaction of Age and Context module

more avatars was significantly higher than those trying to save the avatar on the sidewalk. In comparison to the Quantitative Greater Good module, only minor quantitative differences were found. This shows that the sidewalk altogether has a surprisingly small effect.

Interaction of Age and Context

In the Age-Considering Greater Good and the Influence of Context modules, the influence of age and context was investigated in isolation. The fourth module, was designed to find out whether there was also an interaction of age and context; hence, the city setting with the sidewalk, including child avatars, was used. There were three trials with the following combinations of avatars: two children in the street versus one adult on the sidewalk, one child in the street versus two adults on the sidewalk, and one child in the street versus one adult on the sidewalk.

Results showed that the majority of participants again spared children as opposed to adults, despite the sidewalk context (Fig. 4), as could be expected based on the findings from the previous modules.

In further analyses, two permutation tests were performed to check for differences in the target actions of participants regarding the number of avatars. The conditions with one child in the street and one or two adults on the sidewalk were significantly different from one another ($p < 0.001$). The same held for the comparison of the condition with one child and one adult versus the condition of two children and one adult ($p < 0.05$). The results were in accordance with the

findings from all previous modules. Furthermore, the pattern of the results was compatible with the independent effects of the sidewalk and age.

Self-Sacrifice

The fifth module investigated whether participants value their own life in the VR setup similarly to the value of other avatars. That is, they had the possibility to save avatars at the price of sacrificing their own avatar. In close analogy to the previous modules, participants' choices were investigated, depending on the number of avatars in the group opposing self-sacrifice. In comparison to the non-self-sacrifice condition in the first module, the switch point i.e., the number of avatars in the group necessary to induce consistent decisions, was hypothesized to increase.

The Self-Sacrifice module contained six trials in a mountain setting, where a chasm was implemented on the right lane of the street with a construction sign in front of it. Presented on the left lane were varying numbers of standing adults, ranging from two to seven avatars. The design was created to imply that participants would commit self-sacrifice within the experimental paradigm by driving off the cliff when using the right lane.

Data analysis of this module followed the same procedure as in the Influence of Context module. It was postulated that a fixed threshold could describe the behavior of the participants. In case the number of avatars in the street was below the threshold, the group would be sacrificed. In contrast, when it was above the threshold, participants would choose self-sacrifice. The decisions of only 5.2% of trials were not consistent with such a simple model.

To see if there was a general switching point at which the number of people committing self-sacrifice does not increase anymore, a linear regression was used to fit six models to the data. For these models, the sums of squared residuals were computed and compared. The model with a switch point between the conditions with four and five avatars best described the data (Fig. 5). With a value of 3, the sum of squared errors of this model was much smaller than those of the other models. This indicates that people are consistently willing to sacrifice themselves in the case of being able to save a group of 5 or more avatars with this decision.

Discussion

When driving a car in VR, participants act in favor of the quantitative greater good. That is, their behavior consistently aims at sparing as many avatars as possible. This even applies to situations in which participants have to virtually sacrifice their own avatar to save others. Age and context modulate these behavioral patterns. Specifically, the probability to sacrifice an avatar and its expected remaining lifespan showed an inverse relation. Participants consistently saved younger avatars as opposed to older ones. Surprisingly, the context of sidewalk versus street only had a small influence. In conclusion, the results throughout all conditions support the

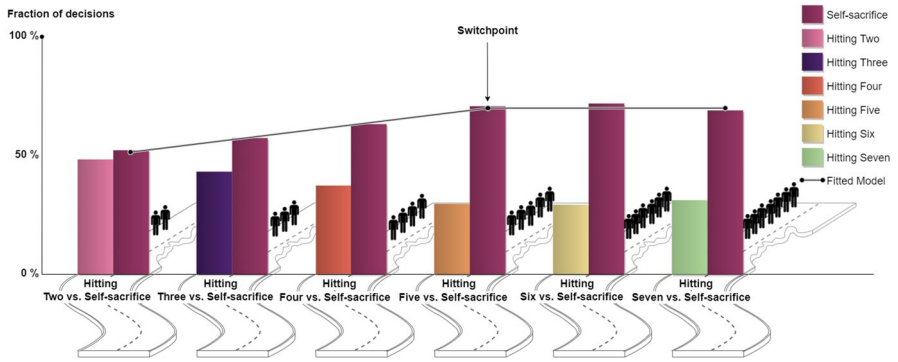

Fig. 5 Decision distribution in the Self-Sacrifice module. Depiction of the best-fitted model for these decisions

hypothesis that people act in favor of the quantitative greater good, even in scenarios involving a sidewalk or self-sacrifice.

Concerning the results of the quantitative greater good module, it could be argued that the participants' decisions were guided by self-preservation, based on a higher expected probability for themselves to survive when hitting only one avatar compared to a group of avatars. However, results of the Self-Sacrifice module indicate that for most participants the influence of self-preservation is limited as already more than half of the participants are willing to commit self-sacrifice for only a group of two avatars. This led us to believe that the decision process was mainly guided by the aim to spare as many avatars as possible.

On the other hand, even though self-preservation seems to be limited, the results show that a considerable portion of participants value the life of their own avatar higher than the life of another avatar as reflected by the existence of individual switch points for avatar groups higher than two. The overall behavior in the self-sacrifice module could have also been influenced by the case that it was not self-evident that hitting one or more avatars would automatically lead to their death. Indeed, after the experiment some participants reported that the car was too slow to kill a human being. Before killing themselves by driving into a bottomless chasm in the self-sacrifice trials, the participants chose to drive into the group of avatars, risking injuries but not lethal damage. Although phrasing the trolley dilemma not in terms of life and death but in terms of health or injury is equivalent, it might lead to differences in decision-making. This should be taken into account when interpreting the results of the Self-Sacrifice module where injuring avatars is opposed to killing their own avatar.

Other complex processes can be assumed to underlie decisions in the age considering greater good module. In social sciences, the term Disability-Adjusted Life Year (DALY) is used (Murray 1994). This is a complex measure that can be roughly understood as the number of years lost in a healthy life. Such a description naturally explains the inverse relation between age and the probability of being spared. Thus, the decision process might be better described not by simply counting the number of

lives, but with a more complex measure, such as the DALY. The possible influence of measures like the DALY can be underlined by the control condition using kneeling adults instead of children, showing that the effect of sparing children in comparison to adults was not purely determined by the size of the avatar's visual appearance. Nevertheless, the aspect that a collision might not necessarily lead to the death of the victim has to be considered when comparing avatars of different ages, too. Besides age and expected lifespan, the likeliness to die in the case of a crash might have influenced the decision of participants to hit adults more often than children. This could also explain the preference to save kneeling adults in comparison to standing adults due to the increased risk of fatal injury when hit by a car in a kneeling position. Hence, it is possible that participants were pondering complex decision processes, including severity of injury and risk of death. These complex decision processes were, contrary to expectations, not substantially affected by a sidewalk, given that participants did not seem to be reluctant to drive on the sidewalk to spare a number of avatars higher than two.

In general, though, decision processes could have been influenced by a lack of incentives. Although VR provides a more realistic experience than surveys or pictorial descriptions, it can still, especially regarding self-sacrifice, be argued that participants in contrast to real-world situations did not have to fear any consequence for their actions. However, it is not possible to implement an experimental feature that represents self-sacrifice or killing other people in a realistic and ethically acceptable way. It is, therefore, questionable whether people would actually be willing to sacrifice themselves in order to act in favor of the quantitative greater good when it comes to such a dilemma situation in traffic or whether their behavior only emerged out of social desirability. This leads to another important aspect, namely that whatever is socially desirable might considerably differ between societies. It is to be noted that since the majority of participants in the current study are Germans and the moral behavior of people with different socio-economic and political backgrounds can significantly vary, the results cannot easily be transferred to other societies.

Moreover, the behavior of participants could have been affected by limitations of graphical display and therefore immersion. This contrasts with some participants dropping out of the experiment because they did not feel comfortable with hitting or even killing virtual avatars. The latter observation does not support a lack of realism or immersion. However, there seem to be many individual differences in play. In this regard, it cannot be ruled out for sure that some participants, especially young ones, were not as committed to the study as expected but were mainly interested in the new VR technology offering a game-like experience. Thus, the average degree of immersion was high, but individual variations should be taken into consideration in future research addressing these problems.

In the field of implementing autonomous driving behavior, empirical knowledge is relatively sparse and ethical approaches are widely debated. Usable ADVs as well as advanced simulation techniques, like 3D VR, are relatively new. Consequently, empirical studies rely heavily on questionnaires directing issues straight at potential customers. The behavior of ADVs and their control algorithms will be judged by the standards and ethics of the societies in which they operate. This again emphasizes the crucial role of acceptance, because self-driving cars need moral algorithms

capable of expressing three aspects: being consistent, not causing public outrage, and not discouraging potential buyers (Bonnefon et al. 2015). For example, the Head of Active Safety of Mercedes Benz, Christoph von Hugo, stated that Mercedes would only build ADVs that would consequently save the driver of the vehicle in hope that this would make the car more attractive to buyers (Morris 2016). But the morality of such automated vehicles should be questioned as there is a gap between what people state how they want an ADV to behave and what they would actually buy (Bonnefon et al. 2016). It is debatable whether ADVs should be programmed based on economic reasons instead of human behavior or ethical arguments.

As mentioned before, traffic is a complex interaction between many road users. The choice to always save one's the own life in a critical situation affects the future response of other road users. There may be a lot of people who would choose a self-sacrificing ADV, but if the majority uses a self-preserving car, this will change. In cases like these, the result could be dramatic for society since the chance of being killed in traffic would rise (Gogoll and Müller 2017). In the following lines, an argument for an obligatory ethics setting in ADVs will be developed. It will also be explain, why a modified trolley dilemma, like it is used in this study, is suitable for a foundation of such a regulated ethics setting.

The abovementioned standpoint of Mercedes Benz represents a moral egoist standpoint. Such a standpoint is plausible if the driver cannot be sure how another car will react in the case of a crash. If the passenger is not disposed to act in favor of the greater good, why take the chance of being killed by a stranger who might act selfishly? This would lead to a prisoner's dilemma like situation. Each road user could choose between self-sacrifice as an analogy for cooperation in the prisoner's dilemma, or self-preservation as an analogy for defection. To maximize the good for the society, it would be adequate to choose a possible self-sacrifice. This would be the lowest toll for society, like in the prisoner's dilemma the lowest combined sentences. This could be an ADV acting in a utilitarian way, as it sacrifices the passengers for the greater good no matter how another road user would act in a crash situation. But if one road user has the possibility to stay alive while the other road user sacrifices her- or himself, the result would likely be that most people prefer the ethics setting of self-preservation to prevent being exploited by people driving, for example, a selfishly programmed vehicle. Hence, even if an individual would like his or her ADV to act in a utilitarian way in traffic, the outcome for society would be worse due to the clash of different ethical settings. For example, an ADV with a strong passenger preservation setting might push a schoolbus full of children into an abyss to ensure the safety of its own passenger. Such a tragedy could be avoided by the same ethical setting for all ADVs, like a utilitarian one. Therefore, one could argue for an obligatory ethics setting, a form of governmental intervention as a common standard for the behavior of ADVs in crash situations. As Gogoll and Müller (2017) showed, a summarized maxim for such a standard could be to minimize harm for all the people involved.

Moreover, this argument also does not allow mixed traffic as an interim solution since this comes close to the situation of personal ethics settings which would lead to the aforementioned prisoner's dilemma, too. A potential solution to this dilemma resides in a contractarian stance. The present findings of participants

acting in a utilitarian way would, however, probably not last long if transferred to a real traffic scenario with a variety of road users and an according variety of intentions in possible crash scenarios. In this case, the observed behavior of participants corroborates the results of the thought experiment by Gogoll and Müller (2017).

In the present study, participants' behavior could be described by utilitarianism more than self-preservation. Yet, not only saving the passenger of the vehicle at all times but also programming a vehicle to behave in utilitarian ways, even in dilemma situations not including self-sacrifice, is contradicted by German law. The first paragraph of the first article of the German constitution states that the human dignity is inviolable, which implies that all humans are equal and contradicts any evaluation of human life. This forbids the saving of one human at the expense of others, as well as any quantifying perspective on human lives. Even if possible solutions for the implementation of completely autonomous cars are contradicted by the law, it is important to discuss these issues in order to advance towards a legal solution in the future. According to Morris (2016), the statement of Mercedes Benz was later revised due to the mentioned legal issues. Nevertheless, ADVs not only have to embody the laws but also the ethical principles of the society they operate in (Gerdes and Thornton 2015).

Regarding the issue of self-sacrifice, it is arguable whether the trolley dilemma is an adequate setting for the moral problems of ADVs. As Gogoll and Müller (2017) state, the trolley dilemma is missing strategic interaction and iteration, meaning that the participants' actions alone determine the result of the dilemma situation. The participant's decision is independent of other human actions. The decision does not take responses of other road users into account. Simply spoken: The situation in a real-life trolley-like scenario is not only determined by a personal ethics setting, but by all involved road users and their ethics settings. The goal of this study was, however, to develop a derived ethics setting which most people agree on. The only way to be able to see what people agree on, even if it is only due to social acceptance, is to isolate the strategic intentions and iterations from the possible scenario. In these cases, the participant's decision is independent of the reaction of different road users as well as of social acceptability since only the participant was able to see the result of her or his action.

The present findings suggest that the gap between a set of mandatory ethical rules, which solve the prisoner's dilemma and the behavior of the participants is not as big as one would expect.

Despite the case that various studies were conducted under the assumption that people would like ADVs to behave similarly to humans (Goodall 2014; Malle et al. 2016) concerns could be raised about the uniform behavior of ADVs. Sikkenk and Terken (2015) found that many factors dramatically influence human behavior in traffic, e.g. weather conditions and the driving style of other traffic participants. Variation does not only occur in driving behavior but also in judging decisions of humans in contrast to those of machines (Malle et al. 2016). This was also shown by Li et al. (2016) who examined the differences in responsibility between humans and machines in cases of inevitable fatal crashes. Participants had to judge the decision of either a human driver or an autonomous car in a dilemma. In contrast to human

drivers, where utilitarian decisions were most favorable, participants expected ADVs to behave in a utilitarian manner under all circumstances.

Different studies point out that the general population seems to favor utilitarian decisions (Bonnefon et al. 2015; Li et al. 2016; Malle et al. 2016). This applies even to cases in which drivers have to sacrifice themselves for the greater good (Sachdeva et al. 2015). Such behavior can be understood as an act of maximizing utility (Thomson 1985). Therefore, it is mostly referred to as utilitarian reasoning and decision-making. Ab initio, utilitarian decisions offer themselves as a quantitative treatment and appear to be suitable for ADVs. However, the problem of how to implement ethics in machines, especially in dilemma situations, remains. A recent project at the Bristol Robotic Laboratories (Winfield et al. 2014) tried to implement Asimov's three laws of robotics[2] to show that robots can be ethical, as well as safe. In this experiment, they defined a puck-like robot as a human which moved towards a hole in a table. Another robot, the one with the implemented robotic laws, had the task to save the robot that moved towards the hole. But the experiment showed that there was no such thing as a simple rule, like the first Asimov law of robotics, to save a human life in dilemma situations. In these situations, the robot showed inconsistent behavior when it should rescue two robots moving towards the hole. Sometimes it was able to rescue one, sometimes even both, but in the worst scenario none was saved. To fix that, more rules have to be applied to the initial code. But who is the one to be saved first? Could there be a rule to prioritize one life over another? A simple solution is not in sight but seems to be a crucial aspect in promoting ADVs to potential customers and allowing them to be an integral part of society. Ethical dilemmas do not necessarily have absolute answers, but they do have significant ethical implications for users. Ethicists serving as experts for ethical evaluations of robots are key to solving the question of how ADVs should behave (Millar 2016).

Despite several unresolved issues, there are many ethical arguments for fully autonomous cars. Not only is it likely that they improve mobility for elderly and disabled people, but also reduce crashes and annual fatalities in traffic, ease congestion, allow more work-related activities while driving, and therefore a productivity gain, improve fuel economy, reduce parking issues, and offer transportation to those unable to drive. For example, the US economic benefits could reach around 25 billion dollars per year with only a 10% market penetration. Including high penetration rates, this raises the annual benefit up to 430 billion dollars, which makes ADVs a technology for a better future (Fagnant and Kockelman 2015). The number of avoided fatalities is a sufficient reason to promote ADVs. Therefore, the idea of

[2] The three laws of robotics (Asimov 1950) were created as a part of a science fiction novel by Isaac Asimov as a concrete beginning of possible ethical settings for robots. They are human centered, and easily applicable to ADVs as well.

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

McBride (2016) for partial automation only is decidedly rejected here. Instead, further research addressing open questions should be encouraged. These range from technical issues to ethical and psychological problems, as well as legal aspects, such as responsibility and policy issues.

In summary, the results show that participants consistently behave in utilitarian ways in various dilemma situations. Their decisions were only slightly modulated by context, such as a sidewalk. Furthermore, the effect of age might be subsumed in an utilitarian decision process as well. Even in conditions involving a self-sacrifice, participants' decisions were compatible with a utilitarian strategy. The study describes driver decisions in possible traffic dilemma situations more accurately than a mere survey. In contrast to previous surveys, it is shown that people are inclined to act in a utilitarian way. The majority of participants acted in a way that their behavior would minimize harm to all road users. This maxim could be derived as a mandatory ethics setting in all future ADVs, because a personal ethics setting could result in a prisoner's dilemma situation and more fatalities in traffic. The option of implementing such a setting is more likely to be accepted if people indeed act in the way discovered as opposed to the way they describe their actions. The study provides a basis for an algorithm implementing morals regarding ADVs. It describes how human car drivers would behave in these conditions and what is therefore seen as adequate behavior in general traffic situations.

Materials and Methods

Participants

Two hundred sixteen unpaid subjects participated in the study. Participants were acquired from various venues throughout Osnabrück. Data from 27 participants had to be excluded from the analysis for various reasons: 15 participants did not complete the experiment due to nausea or disagreement with the experimental settings; 12 participants did not pass training (more than six trials). In the end, data from 189 participants served for analysis (62 female, 127 male). They were aged between 18 and 67 years with a mean of 24.32 years.

Stimuli and Design

In the VR environment participants were driving a car on a one-directional track with two lanes. The environmental surroundings varied between five settings. One suburban and two mountain settings consisted of a dual roadway, where the starting lane was randomized for each trial. The two city settings consisted of a one-way street and a drivable sidewalk; starting lane was always the street. The car was driving at a constant speed of 36 km/h, visible to the participants on the car display, and the track length ranged between 180 and 200 m to avoid habituation (Fig. 6).

At the end of each lane, distinct types of male avatars appeared in different combinations, forcing participants into a dilemma-situation. The lane side on

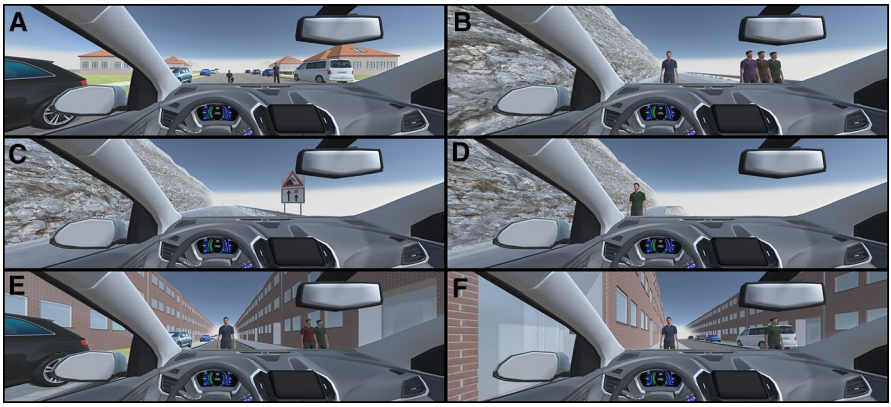

Fig. 6 Screenshots of the VR environment in the different modules. **a** Age-Considering Greater Good module in the suburban setting. **b** Quantitative Greater Good module in the mountain setting 1. **c** Self-Sacrifice module showing the road sign warning of the oncoming chasm in the mountain setting 2. **d** Self-Sacrifice module in the mountain setting 2. **e** Age-Considering Greater Good module in the city setting 2. **f** Age-Considering Greater Good module in the city setting 1

which the two types of avatars appeared was randomized for each trial (excluding modules containing a sidewalk) and did therefore not correlate with the starting lane.

To decrease the visual range and thereby guarantee a constant decision-making time of four seconds, all settings included foggy weather. At the beginning of each trial, a beep indicated to the participants that they had control over the vehicle. The relatively low speed of the car was selected as a compromise to allow reasonable time for deliberation and to have the nature of the obstacle clearly visible. At the same time, it involves the danger that not all participants perceive a car crash as a threat to the life of the avatars. At 15 m from the avatars, another beep signaled that the control over the vehicle was withdrawn, as later inputs would have led to incomplete lane change maneuvers.

Five Hypotheses were tested using different experimental modules.

The *Quantitative Greater Good module* consisted of three trials to test Hypothesis 1. The environment for this module included the suburban and mountain setting 1. There was always one standing adult avatar in one lane (randomized) as opposed to either two, four, or six in the other lane.

The *Age-Considering Greater Good module* aimed at testing Hypothesis 2. It was composed of six trials in the suburban setting. A child, an adult, and an old person were used as avatars. Additionally, a kneeling adult served to make sure that possible effects were not only due to the size of the stimuli, given that the kneeling adult was of the same height as the child. Each trial presented one of the following six combinations of avatars: One child versus one standing adult, one child versus one old person, one standing adult versus one old person, one kneeling adult versus one standing adult, one kneeling adult versus one old person, and one kneeling adult versus one child.

The *Influence of Context module* investigated Hypothesis 3 and consisted of six trials in the city setting with a one-way street plus a driveable sidewalk on the left or right side in the city setting 1 or 2, respectively. The setting was randomized for each trial. There was always one standing adult on the sidewalk and two to seven standing adults in the street, each combination occurring once for every participant.

The *Interaction of Age and Context module* testing Hypothesis 4 was based on the *Influence of Context module* also using the city settings, including children as avatars. There were three trials with the following combination of avatars: two children in the street versus one standing adult on the sidewalk, one child in the street versus two standing adults on the sidewalk, and one child in the street versus one standing adult on the sidewalk.

The *Self-Sacrifice module* investigating Hypothesis 5 contained six trials in the mountain setting 2. Here, a chasm was implemented in the right lane of the street with a construction sign in front of it. The chasm was at the edge of the mountain road and it was bottomless. A vehicle or person falling into the chasm would fall down the steep flank of the mountain. A varying number of standing adults, ranging from two to seven avatars, was presented in the left lane, while the chasm was in the right. It was created to imply that participants would commit self-sacrifice within the experimental paradigm by driving over the chasm when using the right lane.

Procedure

After giving written consent, participants were seated in front of a keyboard and equipped with the Oculus Rift DK2 in combination with Bose Noise-Cancelling Headphones. The VR experiment contained all instructions and consisted of three phases: training trials, experimental trials, and a questionnaire. Participants used the left or right arrow key to change the driving lane. The training phase contained three trials where participants had to avoid three pylons that appeared in one of the lanes alternately. If they hit a pylon, the trial had to be repeated. After successfully completing all three training trials, the various types of avatars were presented to the participants before the experimental trials started. Finally, a questionnaire was answered, which is beyond the scope of the present paper. The duration of the whole experiment was approximately 15–20 min.

Statistical Tests

For all analyses, only final decisions were taken into account.

A permutation test was performed to investige the influences of the number of avatars on participants' decisions in the Quantitative Greater Good, Age-Considering Greater Good and Interaction of Age and Context modules for each condition individually. Additionally, a binomial test using pooled data proved significance in comparison to the null hypothesis of a random distribution of choices in the aforementioned modules.

For the Influence of Context and for the Self-Sacrifice module, fractions of trials that would need to be changed for each participant to show a consistent

decision-behavior were calculated. A fraction of up to 5% could be explained by a natural error rate (Kuss et al. 2005). To test whether the data match the hypothesis of a general switch point, six models of different underlying switch points were fitted to the data and the performance of each was computed. Assuming that upon a certain switch point the number of participants committing self-sacrifice would not further increase, the mean between the conditions with an avatar number higher than a certain switch point was calculated. The model was assumed to pass through this mean in a plateau. For the conditions with avatar numbers smaller than the underlying switch point, a linear increase up to the calculated mean was expected. To test which model fits the data best, the sums of squared residuals were computed and compared.

Acknowledgements The authors would like to thank all study project members: Aalia Nosheen, Max Räuker, Juhee Jang, Simeon Kraev, Carmen Meixner, Lasse T. Bergmann and Larissa Schlicht. This study is complemented by a philosophical study with a broader scope (Larissa Schlicht, Carmen Meixner, Lasse T. Bergmann). The work in this paper was supported by the European Union through the H2020-FETPROACT-2014, SEP-210141273, ID: 641321 socializing sensorimotor contingencies (socSMCs), PK.

Author Contributions This study was planned and conducted in an interdisciplinary study project supervised by Prof. Dr. Peter König, Prof. Dr. Gordon Pipa, and Prof. Dr. Achim Stephan. Maximilian Alexander Wächter, Anja Faulhaber, and Silja Timm shaped the experimental design to a large degree. Leon René Sütfeld had a leading role in the implementation of the VR study design in Unity. Anke Dittmer and Felix Blind contributed to VR implementation. Anke Dittmer, Felix Blind, Silja Timm, and Maximilian Alexander Wächter contributed to the data acquisition, analysis, and writing process. Anja Faulhaber contributed to the data acquisition and the writing process.

Financial Interests This publication presents part of the results of the study project "Moral decisions in the interaction of humans and a car driving assistant". Such study projects are an obligatory component of the master's degree in cognitive science at the University of Osnabrück. It was supervised by Prof. Dr. Peter König, Prof. Dr. Gordon Pipa, and Prof. Dr. Achim Stephan. Funders had no role in the study's design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

References

Asimov, I. (1950). *I, Robot*. Greenwich, CT: Fawcett Publications.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? arXiv:1510.03346.

Bonnefon, J., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1574. https://doi.org/10.1126/science.aaf2654.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice, 77,* 167–181. https://doi.org/10.1016/j.tra.2015.04.003.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5,* 5–15.

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte* (pp. 87–102). Berlin: Springer. https://doi.org/10.1007/978-3-662-45854-9_5.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics, 23*(3), 681–700.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In G. Meyer & S. Beike (Eds.), *Road vehicle automation* (pp. 93–102). New York: Springer. https://doi.org/10.1007/978-3-319-05990-7_9.

Hars, A. (2016). Transformations 2025: How Volkswagen prepares for the (driverless?) future. Resource document. *Driverless-Future*. http://www.driverless-future.com/?p=1019. Accessed November 19, 2017.

Hevelke, A., & Nida-Rümelin, J. (2014). Selbstfahrende Autos und Trolley-Probleme: Zum Aufrechnen von Menschenleben im Falle unausweichlicher Unfälle. *Jahrbuch für Wissenschaft und Ethik, 19*(1), 5–24. https://doi.org/10.1515/jwiet-2015-0103.

Hevelke, A., & Nida-Rümelin, J. (2015a). Ethische Fragen zum Verhalten selbstfahrender Autos. *Zeitschrift Für Philosophische Forschung, 69*(2), 217–224. https://doi.org/10.3196/004433015815493721.

Hevelke, A., & Nida-Rümelin, J. (2015b). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics, 21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision, 5*(5), 478–492. https://doi.org/10.1167/5.5.8.

Li, J., Zhao, X., Cho, M., Ju, W., & Malle, B. (2016). *From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars*. SAE Technical Paper No. 2016-01-0164. https://doi.org/10.4271/2016-01-0164.

Lin, P. (2013). The ethics of autonomous cars. Resource document. *The Atlantic*. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360. Accessed November 19, 2017.

Lin, P. (2015). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte* (pp. 69–85). Berlin: Springer. https://doi.org/10.1007/978-3-662-45854-9_4.

Madary, M., & Metzinger, T. (2016). Real virtuality: A code of ethical conduct. Recommendations for good scientific practice and the consumers of vr-technology. *Frontiers in Robotics and AI, 3,* 3. https://doi.org/10.3389/frobt.2016.00003.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2016). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In McBride, N. The ethics of driverless cars. *ACM SIGCAS Computers and Society, 45*(3), 179–184. https://doi.org/10.1145/2874239.2874265.

McBride, N. (2016). The ethics of driverless cars. *ACM SIGCAS Computers and Society, 45*(3), 179–184.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007.

Millar, J. (2016). An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars. *Applied Artificial Intelligence, 30*(8), 787–809. https://doi.org/10.1080/08839514.2016.1229919.

Morris, D. Z. (2016). Mercedes-Benz's self-driving cars would choose passenger lives over bystanders. Resource document. *Fortune*. http://fortune.com/2016/10/15/mercedes-self-driving-car-ethics/. Accessed November 11, 2017.

Murray, C. J. (1994). Quantifying the burden of disease: The technical basis for disability-adjusted life years. *Bulletin of the World Health Organization, 72*(3), 429–445.

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem". *Emotion, 12*(2), 364–370. https://doi.org/10.1037/a0025561.

Pan, X., Banakou, D., & Slater, M. (2011). Computer based video and virtual environments in the study of the role of emotions in moral behavior. In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Affective computing and intelligent interaction* (pp. 52–61). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-24571-8_6.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*(1), 94–107. https://doi.org/10.1080/17470919.2013.870091.

Sachdeva, S., Iliev, R., Ekhtiari, H., & Dehghani, M. (2015). The role of self-sacrifice in moral dilemmas. *PLoS ONE, 10*(6), e012740. https://doi.org/10.1371/journal.pone.012740.

Sikkenk, M., & Terken, J. (2015). Rules of conduct for autonomous vehicles. In G. Burnett (Ed.), *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications—Automotive UI'15* (pp. 19–22). New York: ACM Press. https://doi.org/10.1145/2799250.2799270.

Skulmowski, A., Bunge, A., Kaspar, K., & Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: A virtual reality and eye tracking study. *Frontiers in Behavioral Neuroscience, 8,* 426. https://doi.org/10.3389/fnbeh.2014.00426.

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based-models and influences of time pressure. *Frontiers in Behavioral Neuroscience, 11,* 122. https://doi.org/10.3389/fnbeh.2017.00122.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217. https://doi.org/10.2307/796133.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.

Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. New York, Oxford: Oxford University Press. https://doi.org/10.1093/0195108590.001.0001.

Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (Eds.), *Advances in autonomous robotics systems. TAROS 2014. Lecture notes in computer science* (Vol. 8717, pp. 85–96). Cham: Springer. https://doi.org/10.1007/978-3-319-10401-0_8.