How should autonomous vehicles behave in moral dilemmas? Human judgments reflect abstract moral principles

Derek Powell[1] (derekpowell@ucla.edu)
Patricia Cheng[1] (cheng@lifesci.ucla.edu)
[1]Department of Psychology, University of California, Los Angeles,
Los Angeles, CA 90095 USA
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Self-driving autonomous vehicles (AVs) have the potential to make the world a safer and cleaner place. A challenge confronting the development of AVs is how these vehicles should behave in traffic situations where harm is unavoidable. It is important that AVs behave in ethically appropriate ways to mitigate harm. Ideally, they should obey a system of principles that both concur with human moral judgments and are ethically defensible. Here we compare people's moral judgments of AV programming with their judgments about the behavior of human drivers, with the goal of beginning to identify such principles. As many debates within ethics remain unresolved, empirical investigations like ours may guide the development of ethical AVs (Bonnefon et al., 2015). In addition, people's judgments about the behavior of AVs may serve as a window into the abstract principles people apply in their moral reasoning.

Keywords: Ethics; Moral Judgment; Robotics

Introduction

A number of auto manufacturers and tech companies are working to develop self-driving autonomous vehicles (AV). These developments hold great promise: creating safer roads (Gao et al., 2014), alleviating traffic congestion (Van Arem et al., 2006), reducing pollution (Spieser et al., 2014), and of course removing the tedious burden of driving through traffic. To be effective, autonomous vehicles must overcome a number of challenges: they must navigate to their destinations, steer properly to remain on roadways, and, perhaps most challengingly, identify objects and predict movements of pedestrians and other vehicles to avoid collisions. These challenges are rapidly being overcome, and experimental AVs like the Google Car have successfully driven for thousands of miles on public roadways (Waldrop, 2015).

Inevitably, AVs will face decisions with moral consequences—situations where an impending collision may injure or even kill human drivers or pedestrians. AVs will have to decide how to mitigate the harm caused by these situations in morally appropriate ways. Though they are by no means easily solved, navigation, steering, and object detection and avoidance are relatively well-defined problems. That is, it is clear what constitutes success: for example, arriving at the correct destination, staying on the correct side of the road, and braking before hitting a pedestrian, respectively. In contrast, moral judgments are sometimes ill-defined: individuals may differ in their moral judgments, and there is considerable disagreement within the field of ethics over how judgments should be made.

Traffic conditions may pose difficult choices similar to the classic Trolley dilemma. In the Trolley or Switch dilemma, a runaway Trolley threatens to kill five men working on a track unless it is redirected toward a side track with only one person working. The situation poses a dilemma: maximizing utility (saving the five) requires violating a moral rule (harming another person). Faced with the Switch dilemma, people generally make utilitarian judgments. The majority of people (85%; Hauser et al., 2007) judge that it is morally acceptable to flip the switch, redirecting the train away from the five and toward the one.

Vehicles in traffic might face dilemmas like the Switch dilemma when mechanical failures or road conditions prevent them from stopping or when other drivers act unpredictably. Should AVs be programmed to sacrifice lives if this produces better consequences overall? Bonnefon et al. (2015) presented participants with a switch-like scenario where an unavoidable deadly collision was about to occur with a group of 10 pedestrians unless an AV turned toward a single pedestrian. Participants approved of the AV turning to kill the one at rates very similar to those for human drivers. Participants were also generally willing to allow AVs to sacrifice the lives of their passenger by swerving to collide with a wall rather than a group of pedestrians. These researchers conclude that people are willing to accept AVs programmed for utilitarian sacrificial behaviors in at least some circumstances.

However, there are some moral situations where people refuse to sacrifice even when it would create the best

outcomes. For instance, although the vast majority of people approve of sacrificing one to save five in the classic Trolley dilemma, in the very similar Footbridge or Push dilemma a majority refuse to intervene. In the Push dilemma, a runaway trolley threatens to kill five workmen on a track. A large man is standing on a footbridge over the track, and the only way to save the five workmen is to push this man off the footbridge so that his body will stop the trolley. Approximately 88% of people refuse to push the man (Hauser et al., 2007). Whereas people are sometimes willing to trade-off between moral rules and better outcomes, at other times they are unwilling to do so—even for comparable outcomes. Will people approve of sacrificial behaviors from AVs in dilemmas that more closely resemble the Push dilemma, in contrast to their judgments about AVs facing switch-like traffic problems (Bonnefon et al., 2015)?

As it is impossible to anticipate all possible traffic scenarios in which an AV might find itself, AVs will need to respond to novel traffic situations according to abstract principles. Human moral judgments do not appear to perfectly adhere to any simple normative proposal, and researchers have offered divergent accounts of the principles underlying their judgments. For instance, some have argued that differences in Push and Switch judgments owe to affective reactions to the use of personal force (e.g. Greene et al., 2001) whereas others have argued that these judgments are the result of sophisticated deontological rules like the "doctrine of double effect" (DDE), which prohibits intentionally harmful actions, but may permit actions that produce a greater good for which harm is a foreseen but unintended consequence (Mikhail, 2011; Hauser et al., 2007). It is unclear how these proposals would apply to programmed AVs—it doesn't seem as if such programming involves "personal force" and the role of intentions is rather murky in this context. Can AVs even be said to have intentions in the sense of other moral agents?

Nevertheless, the development of moral AVs will require us to craft a system of principles that both satisfies human moral judgments in most cases and that is ethically defensible—even if imperfect.

Waldmann and colleagues (Waldmann & Dieterich, 2007; Wiegmann & Waldmann, 2014) have proposed that the causal structure of a moral situation is an important determinant of people's moral judgments. Their theory is related to the DDE, which claims that intentionality is inferred on the basis of the causal structure underlying the moral dilemma. However, rather than the intention of the agent, in Waldmann et al's theory it is the causal structure itself that drives moral intuitions. According to their theory, the locus of interventions in a causal system influences the attentional focus to different aspects of the moral dilemma, which in turn affects moral intuitions. In the switch dilemma, flipping the switch acts as a common cause of both the killing of the one on the side track and the saving of the five on the main track, highlighting both of these outcomes. In contrast, pushing the man off the bridge in order to stop the trolley represents a causal chain structure, and focuses subjects initially on the fate of the one victim.

This account leaves open the question of *why* causal models should be morally relevant. One possible answer, suggested by Kamm (2015), is built on the assumption that victims are endowed with rights (e.g., the right not to be harmed) and causal models highlight inter-victim relations. Victims stand in a *substitutive* relation if their roles in a situation could be arbitrarily swapped as, for instance, in the common-cause scenario. Shepard (2008) calls a similar concept the *symmetry* principle of *invariance under permutation of individuals*. Shepard regards this principle as a necessary overarching constraint for moral acts: An act is morally acceptable to such an extent as it would remain acceptable if individuals in a situation were permutated into different roles. In the trolley dilemma, sacrificing one to save five satisfies this criterion.

In contrast, in a causal chain scenario such as the footbridge, victims are in what Kamm (2015) terms a *subordinative* relation. Cause and effect, unlike effects of a common cause, are asymmetrical and not arbitrarily substitutable. Here Kamm argues that it violates our understanding of human rights to harm a person as a means of a later occurring greater good. Thus, the substitutability of agents' roles in the common-cause scenario, but not the causal chain scenario, could provide a justification for the seemingly opposite moral intuitions in a Switch and a Push dilemma.

The importance of causal structure (or likewise, of symmetry or substitutability) suggests that people's moral judgments may be influenced by the level of abstraction or concreteness at which they consider a situation. The Switch case represents an extreme—a rare case of perfect symmetry. In contrast, for situations with more complex causal structures—essentially any case that is not a simple artificial dilemma—there are many avenues by which varying degrees of asymmetry might be produced. If the intervention in the Switch scenario is considered concretely, the perfect symmetry of the scenario licenses the sacrifice of the one to save the five. However, if we consider the intervention in the abstract, other scenarios would be possible, and few of these will have perfectly symmetrical causal structures. On this account, encouraging abstract consideration of the dilemma may reduce approval for sacrificial actions. Making moral judgments about the programming of autonomous vehicles should encourage this sort of abstraction by introducing the possibility of other causal scenarios in which this programming will be applied. Thus, in considering moral judgments at a greater level of abstraction, we might expect people to less strongly endorse

sacrificial actions by AVs than in situations in which a human drives the vehicle.

The present experiments

To investigate how evaluating moral dilemmas at differing levels of abstraction would affect moral judgments, our experiments examined people's moral judgments of specific moral instances involving human drivers as compared with judgments about how AVs should be programmed to behave in these and other similar situations. In addition, the wording of the instructions were manipulated to encourage concrete or abstract consideration of the cases. We predicted that people would less strongly approve of sacrificial actions by programmed AVs than by human drivers, and when they are encouraged to think abstractly, in common-cause scenarios due to the possibility of the application of this judgment in other causal scenarios. However, no differences in judgments were predicted between judgments of AVs and human drivers in causal chain scenarios because these scenarios are already strongly aversive due to the asymmetrical inter-victim relation implied by the chain structure.

A secondary goal of the study was to examine people's moral judgments about AV programming in moral dilemmas with varied causal structures. Bonnefon et al. (2015) found that people are at least sometimes willing to allow AVs to sacrifice human lives to save the lives of others. However, their study only examined trolley-like cases with a common-cause structure, leaving open questions about the generality of their findings.

Experiment 1

Experiment 1 examined participants' moral judgments for dilemmas where a vehicle, either driven by a human or a driverless AV steered by a computer program, faced an unavoidable and deadly collision. We compared a car analog of the Switch and Push dilemma. In addition to manipulating the type of vehicle and dilemma, we also manipulated the abstractness of the problem description and the phrasing of the judgment probe for the AV condition, resulting in a concrete AV condition and an abstract AV condition. The main goal of the experiment was to study how abstractness of the steering mechanism (a driver facing a specific situation vs. a general program for the AV) would affect intuitions about Switch and Push cases.

Design. The experiment followed a 2 x 3 factorial between-subjects design (dilemma x abstraction): Participants were assigned either the Switch or Push dilemma, and to one of 3 levels of abstraction: either the human driver condition, the concrete AV condition, or the abstract AV condition.

Participants. Participants were 413 workers (237 female, median age = 32 years old) recruited from Amazon's Mechanical Turk (mTurk) work distribution website. Workers were paid $0.25 to participate in the study.

Materials. The Switch dilemma described a situation where the brakes of a truck have failed and the truck is headed toward a red car with three passengers. The driver (human condition) or the computer system directing the truck (AV conditions) must choose whether to continue on their present course, killing the three passengers in the red car, or to turn into a yellow car waiting at an intersection on a side street with only one passenger. Unfortunately, the truck cannot safely turn off the road onto a sidewalk as they are all full of pedestrians who will also be killed.

In the Push dilemmas, a runaway truck is again headed toward a red car with three passengers. However, in this dilemma the vehicle of interest is a blue car waiting at an intersection behind a yellow car with one passenger. The driver (human condition) or the computer directing the blue car (AV conditions) must decide whether to push the waiting yellow car with one passenger from the side street into the path of the runaway truck, saving the three passengers in the red car but killing the one in the pushed car. Each dilemma was accompanied by a diagram that depicted the situation, an example of which is shown in Figure 1.

The level of abstraction was manipulated between the concrete AV and abstract AV conditions by including additional wording in the abstract AV condition highlighting the generality of the principles employed by the AV:

"… In fact, there are thousands of possible dangerous traffic situations whose precise characteristics one cannot possibly anticipate. It is therefore crucial that the driverless car be equipped with a computer program that implements general rules designed to be applicable across a large variety

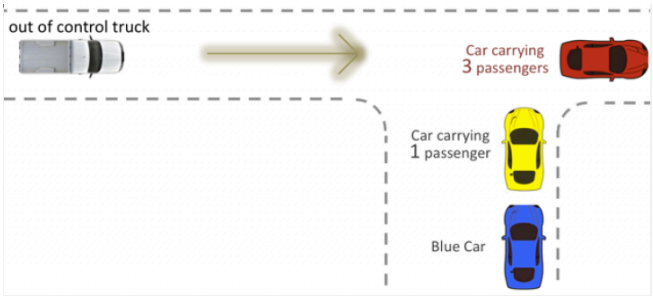

Figure 1: Diagram of traffic situation that was shown to participants for the Push dilemmas. The blue car is either an autonomous vehicle or is driven by a human driver, and must choose whether to push the yellow car into the path of the truck.

of possible scenarios. The program then directs how the car should best behave in these critical situations."

After reading the dilemmas participants were asked to make a moral judgment. In the human conditions participants were asked "Would it be appropriate for the [truck to turn onto the side street / blue car to push the yellow car into the crossing]?" for Switch and Push dilemmas, respectively. In the AV conditions participants were asked, "Would it be appropriate to design the steering function so that the [driverless truck turns onto the side street / blue car pushes the yellow car into the crossing]?"

The level of abstraction between the concrete AV and abstract AV conditions was also manipulated by additional language introducing the questions and problems. Before the "would it be appropriate …" question was asked, the abstract AV condition added instructions to "Imagine a world in which the steering function of a driverless car would decide that the blue car should push the yellow car into the crossing."

Procedure. Participants were recruited from mTurk and redirected to a Qualtrics survey website where study procedures were administered. Participants gave their consent to participate and answered some brief demographic questions before they were randomly assigned to a condition. Participants were then given some context about the topic of the study. Those assigned to the AV conditions read a brief explanation about the development of driverless cars, wherein it was explained that the cars would sometimes have to make decisions where an accident was unavoidable. Participants in the human driver conditions read a similarly worded introduction, but with the discussion of driverless cars omitted. Participants then considered their assigned dilemma and made a moral judgment about whether it was appropriate for the driver to take action in the dilemma. These judgments were made on a Likert scale ranging from 1 (completely appropriate) to 6 (completely inappropriate). Finally, participants answered some brief questions about how they made their judgments and whether they had ever seen the dilemmas before, two simple comprehension check questions (e.g., "two plus two is equal to what?") and whether they had paid attention and taken their participation seriously.

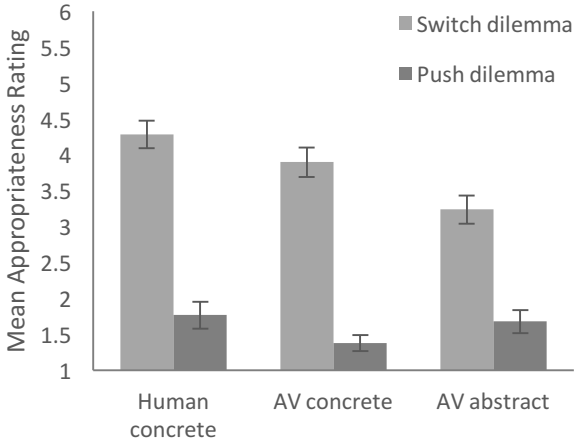

Figure 2: Participants' moral judgments in Experiment 1.

Results. Of the 413 participants recruited, 99 failed at least one comprehension check or indicated they had not paid attention. These participants were excluded, leaving 314 participants in the following analyses.

Participants' moral judgments are shown in Figure 2. These judgments were examined with a 2 x 3 (dilemma x abstraction) between-subjects ANOVA. Participants approved of acting in the Switch cases much more strongly than in the Push cases for all conditions, as indicated by a significant main effect of dilemma, $F(1, 308) = 231.5$, $p < .001$.

In support of our causal model explanation of the Switch and Push dilemmas, the main effect of abstraction was significant ($F(2, 308) = 5.359$, $p = .005$) and there was a significant interaction, $F(2, 308) = 4.898$, $p = .008$. The interaction appears to be driven by a difference between conditions for the Switch dilemma, and an absence of differences for the Push dilemma. Whereas significant differences were observed for the Switch dilemma between the abstract AV condition and the human condition ($t(103) = 3.886$, $p < .001$), as well as between the abstract AV condition and the concrete AV condition ($t(104) = 2.355$, $p = .02$), no significant differences were observed among the Push scenarios (all ps $> .05$). The contrast between the concrete AV and human condition was also non-significant, $t(101) = 1.39$, $p = .168$.

Experiment 2

To further test the effect of abstractness, we added an abstract human driver condition in Experiment 2. In other words, we added a manipulation of abstraction by wording. Two additional goals are to test the replicability of the results of the previous study and to evaluate a floor effect as an alternative explanation of the lack of influence of driver (human vs AV) for the Push dilemma.

Participants. Participants were 610 workers (351 female, median age = 32 years old) recruited from Amazon's Mechanical Turk (mTurk) work distribution website. Workers were paid $0.25 to participate in the study.

Design, Materials, and Procedure. The materials and procedure of this experiment were nearly identical to those of Experiment 1 save for the addition of an abstract human driver condition for Switch and Push dilemmas. This resulted in a 2 x 2 x 2 factorial design (dilemma x driver x wording).

In the abstract human driver conditions, participants were introduced to the dilemmas in the context of developing a training manual for delivery and ridesharing companies. The introduction was similar to the abstract AV conditions, stressing that the rules must be made to apply across many different scenarios. For the Switch and Push variants of these conditions, the moral judgment question read "Imagine a world where general guidelines in the transportation company manual prescribe that the driver of the [truck should turn onto the side street and hit the yellow car / blue car should push the yellow car into the crossing]. Would it be appropriate to write the general guidelines so that a company driver [would turn the truck onto the side street / in the blue car would push the yellow car into the crossing]?"

To examine the possibility of floor effects for the Push items, participants who were assigned to the Push dilemma conditions were also assigned to make a judgment about a Transplant dilemma. This dilemma asks participants to judge whether it is acceptable for a doctor to kill a patient in order to use his organs to save several other patients.

Results. Of the 610 participants originally recruited, 83 were excluded for failing at least one comprehension check or for indicating that they had not paid attention, leaving 527 participants in the final analysis.

Participants' moral judgments in Experiment 2 are shown in Table 1. These judgments were analyzed using a 2 x 2 x 2 (dilemma x driver x wording) between-subjects ANOVA. As in Experiment 1, approval was much lower for Push dilemmas than Switch dilemmas, indicated by a significant effect of dilemma, $F(1, 519) = 422.8$, $p < .001$. Replicating the results of Experiment 1, moral approval was lower in AV conditions than human conditions for the Switch dilemma but not the Push dilemma, as indicated by the two-way interaction between dilemma (switch or push) and driver (human or AV), $F(1, 519) = 4.351$, $p = .037$.

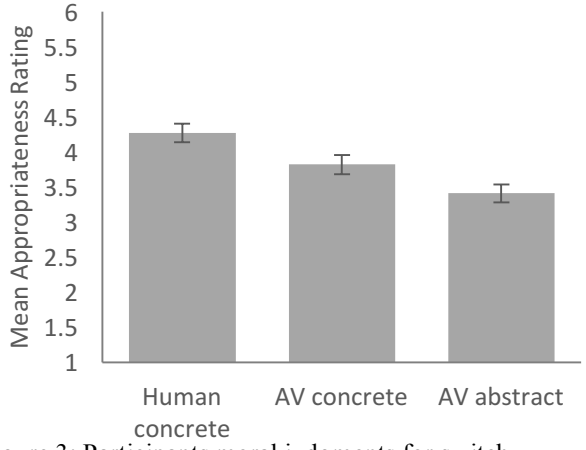

Figure 3: Participants moral judgments for switch dilemmas averaged from Experiments 1 and 2.

Also as predicted, moral approval generally appears to be lower for abstract wordings than for concrete wordings, although this effect was not significant, $F(1, 519) = 2.889$, $p = .09$. Although this manipulation was meant to introduce a greater level of abstraction in the same way as the AV conditions, the manipulation was unfortunately not fully equated. For example, the manual explicitly reminded subjects of the specifics of the dilemmas (Switch, Push), whereas the instructions for abstract AV condition did not. This may have made it more difficult for participants in the abstract human condition to invoke abstract principles.

Table 1. *Mean moral appropriateness judgments by dilemma (switch, push), driver (human, AV) and wording (concrete, abstract) from participants in Experiment 2.*

		Human	AV
Switch	Concrete	4.25 (1.49)	3.75 (1.55)
	Abstract	3.83 (1.65)	3.54 (1.48)
Push	Concrete	1.54 (1.18)	1.45 (1.04)
	Abstract	1.29 (0.85)	1.55 (1.14)

Note: Means with standard deviations in parentheses.

A further investigation of participants' Switch dilemma judgments in a 2 x 2 ANOVA (driver x wording) reveals a significant effect of driver ($F(1, 260) = 4.348$, $p = .038$). These findings qualitatively replicate the differences between conditions in Experiment 1, although only the contrast between the abstract AV and concrete human condition was significant, $t(122) = 11.36$, $p < .001$. All other effects were non-significant (all Ps > .06).

The absence of differences among the Push scenarios seems unlikely to be a simple floor effect, as indicated by the still lower approval for the Transplant dilemma (Mean = 1.29, SD = .819), $t(262) = 2.771$, $p = .006$.

Meta-Analysis of Experiments 1 and 2

With the exception of abstract human driver conditions, the materials and procedures of Experiments 1 and 2 are identical, allowing data from these experiments to be pooled for their shared conditions for increased statistical power.

Of particular interest are comparisons between the human, concrete AV and abstract AV conditions for the Switch dilemma, as qualitatively similar yet somewhat different patterns of results were observed in Experiments 1 and 2. Pooling these data reveals significant contrasts for all pairwise comparisons: lower approval was observed for the abstract AV condition as compared with the concrete AV ($t(240) = 2.125$, $p = .035$) and human conditions, $t(238) = 4.609$, $p < .001$. Lower approval was also observed for the concrete AV condition as compared with the human condition, $t(228) = 2.329$, $p = .021$.

Discussion

Across two experiments we found that people were generally willing to allow AVs to act to sacrifice one life in order to save three others in a common-cause scenario like the Switch but not in a causal chain scenario like the Push. However, as compared with judgments of human drivers, people were less willing to see AVs sacrifice in a common-cause dilemma. There was no analogous difference for a causal-chain dilemma. These findings are consistent with the role of causal structure in human moral judgments and with the constraint of symmetry based on Kamm's (2015) analysis of inter-victim relationships (see also Waldmann, Wiegman, & Nagel, in press).

Our main goal was to provide a theoretical account that both concurs with human moral judgments and is ethically defensible. We tested and confirmed the predicted interaction between the abstractness of construals of moral dilemmas and the causal structure of moral dilemmas. We acknowledge, however, that the experiments were not designed to test our account of the interaction against alternatives. It may be possible, for example, to derive predictions from two-system theories, although we do not see how they can make plausible predictions about abstractness. If anything, this account should seem to predict more utilitarian reasoning in abstract cases, which should not trigger emotional involvement, contrary to what we found.

Another factor that is often neglected in research on moral dilemmas concerns legal considerations. Trolley dilemmas describe rare situations which do not routinely happen. Therefore, there are no clear legal regulations about such accidents. By contrast, accidents involving cars are frequent. Traffic is strictly regulated by laws. In fact, articles about AVs typically focus on the possibility of accidents and on issues of liability. Thus, it seems plausible to assume that a programmer of an AV steering mechanism will try to prevent situations in which the AV either kills its owner or innocent bystanders. However, legal considerations seem less able to explain the lower moral approval we observed in the abstract AV condition compared to the concrete AV condition, both in Experiment 1 and in our meta-analysis. Both conditions involved the program in AVs operating across all traffic situations, and the company developing the program would be equally liable.

The development of ethical autonomous machines is an important and exciting application of the budding field of experimental ethics, to which we hope significant further inquiry will be devoted. Keeping pace with the rapid development of AI research and engineering certainly demands it.

References

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

Gao, P., Hensley, R., & Zielke, A. (2014). A road map to the future for the auto industry. *McKinsey Quarterly*.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–8.

Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*, 1–21.

Kamm, F. M. (2015). *The trolley problem mysteries*. New York: Oxford University Press.

Shepard, R. (2008). The step to rationality: The efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, *32*, 3-35.

Spieser, K. et al (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In G. Meyer and S. Beiker (Eds.), *Road Vehicle Automation*, 229–245. Springer.

Van Arem, B., Van Driel, C. J., & Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Transactions on Intelligent Transportation Systems*, *7*, 429– 436.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychological Science*, *18*, 247–53.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (in press). Causal models mediate moral inferences. In J.-F. Bonnefon & B. Trémolière (Eds), *Moral inferences*. Hove: Psychology Press.

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature*, *518*, 20–23.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*(1), 28–43.