AI & SOCIETY
https://doi.org/10.1007/s00146-023-01813-y

OPEN FORUM

Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicles

Dario Cecchini[1] · Sean Brantley[1] · Veljko Dubljević[1]

Received: 3 May 2023 / Accepted: 30 October 2023

Abstract

The imminent deployment of autonomous vehicles requires algorithms capable of making moral decisions in relevant traffic situations. Some scholars in the ethics of autonomous vehicles hope to align such intelligent systems with human moral judgment. For this purpose, studies like the Moral Machine Experiment have collected data about human decision-making in trolley-like traffic dilemmas. This paper first argues that the trolley dilemma is an inadequate experimental paradigm for investigating traffic moral judgments because it does not include agents' character-based considerations and is incapable of facilitating the investigation of low-stakes mundane traffic scenarios. In light of the limitations of the trolley paradigm, this paper presents an alternative experimental framework that addresses these issues. The proposed solution combines the creation of mundane traffic moral scenarios using virtual reality and the Agent-Deed-Consequences (ADC) model of moral judgment as a moral-psychological framework. This paradigm shift potentially increases the ecological validity of future studies by providing more realism and incorporating character considerations into traffic actions.

Keywords Traffic moral judgment · Moral machine experiment · Trolley problem · ADC model

1 Introduction

Autonomous vehicles (AVs) are expected to radically change transportation by bringing many benefits like decreased traffic accidents, pollution reduction, and economic growth (Dubljević et al. 2021). Many vehicle manufacturers are investing in this transition, promising the production of fully automated cars, freight trucks, and buses in the near future. However, some moral philosophers warn that the current traffic regulations are inadequate to tackle some ethical issues arising from the development of AVs, such as privacy, cybersecurity, and accountability in the case of an accident (Lütge et al. 2021). The hope of some scholars is that AVs will be programmed to make suitable moral decisions in relevant traffic situations (Millar 2017; Lin 2016).

Since different ethical views may disagree on how to inform AV ethical settings, it seems reasonable to study people's moral preferences in traffic situations. Indeed, this was the goal of the moral machine experiment (MME), a highly influential online study, which, thanks to its simple and repeatable experimental design, collected millions of responses about unavoidable accidents (Awad et al. 2018). However, the MME has many limitations (Harris 2020; Cunneen et al. 2020; Geisslinger et al. 2021; Etienne 2022). In particular, the employment of high-stakes sacrificial dilemmas inspired by the trolley thought experiment is problematic because they are simplistically binary and lack ecological validity (Bauman et al. 2014). Other recent studies also evaluating moral traffic decisions were able to obtain more ecological validity thanks to the use of virtual reality (Sütfeld et al. 2017; Faulhaber et al. 2019; Li et al. 2019; Grasso et al. 2020), although they still rely on trolley-like cases. As we will argue, this experimental paradigm is particularly inadequate to inform AVs because it does not include agents' character-based considerations in the dilemmas and is incapable of facilitating the investigation of low-stakes mundane traffic scenarios (Himmelreich 2018).

A paradigm shift in the study of traffic moral judgments is needed to inform future research in AV ethics and eventually to create aligned software. Our original solution combines the creation of mundane traffic moral scenarios using virtual reality and the agent-deed-consequence (ADC) model of moral judgment as a moral-psychological framework

✉ Veljko Dubljević
veljko_dubljevic@ncsu.edu

[1] Department of Philosophy and Religious Studies, North Carolina State University, Raleigh, NC, USA

Published online: 29 November 2023

(Dubljević and Racine 2014; Dubljević 2020; Dubljević et al. 2018). We will show how all such elements merged together can increase the ecological validity of future studies and go beyond the trolley paradigm.

This paper proceeds as follows. First, we briefly establish the necessity of ethically acting AVs and the importance of obtaining consistent human moral preferences about AVs to reach that purpose (Sect. 2). We highlight the main flaws of the MME and the trolley paradigm in Sect. 3. Then, in Sect. 4, we outline the main experimental benefits brought by virtual reality technology and some reasons to study mundane traffic situations. In Sect. 5, we introduce the ADC model of moral judgment and discuss its translation into the traffic context. Finally, in Sect. 6, we describe our experimental design and explain its main advantages over the trolley paradigm, as well as its current limitations.

2 Background: the ethical challenge of autonomous vehicles

An AV is a vehicle that can fully or partially drive without a human operator for a prolonged time. The Society of Automotive Engineers (SAE) identifies six levels of automation: from level 0 (no automation), which corresponds to traditional vehicles, to level 5 (full automation), in which the vehicle can perform all driving functions without human assistance. Although only partial automation vehicles (level 2–level 3), like *Tesla S*, have been commercialized to date,[1] the deployment of higher levels of automated vehicles in the future is likely. High automation vehicles (level 4), such as the *Google Waymo* project, are currently being tested, and companies, such as General Motors and Apple, have announced that fully AVs (level 5) will be in production by 2026 (Hawkins 2022; MacRumors 2023). As Goodall (2018) states, crucial for AVs' full automation is the capacity to detect objects, classify them, and predict their movements, speed, and directions. Although current AV sensors fail to accomplish the last two tasks,[2] these advancements are expected within the decade.

According to some scholars, deploying high or fully automated vehicles raises important ethical issues and the need to develop an ethically informed AV, that is, an AV that responds to environmental stimuli by following previously agreed-upon *ethical settings* (Millar 2017; Lin 2016; Lütge et al. 2021). The AVs software architecture has three layers structured similarly to a nervous system; a perceptual layer that utilizes sensory equipment, a planning layer composed of search spaces and planning algorithms, and a trajectory layer to guide movement (Sharma et al. 2021). Ethical settings would be embedded in the vehicle's planning layer, permitting it to make ethical decisions in relevant situations (Millar 2017).

Arguably, programming AVs with ethical settings is necessary for two primary reasons. First, since accidents with AVs are unavoidable, AVs will have to make functional equivalents of moral decisions. Second, if AVs are to be part of the traffic community, private consumers will need to trust their autonomous decisions.

An estimated 94% of traffic accidents were attributed to human error (Singh 2018). Thus, the implementation of AVs will have the potential benefit of reducing the considerable harm caused by car crashes.[3] Yet, accidents with AVs do still happen.[4] To the extent that AVs have limited maneuverability, they may not have sufficient time to avoid collisions with objects that suddenly change direction (Goodall 2014). Furthermore, the probability of an accident increases in the likely scenario in which AVs and non-automated vehicles coexist. Importantly, in the case of an unavoidable collision, an AV will have to choose between two or more unfavorable alternatives. For example, it has been highlighted that AV developers and legislators must decide whether to protect AV passengers at the cost of endangering pedestrians or prioritize pedestrians by risking losing AV consumers' trust (Bonnefon et al. 2016). Another potential dilemma arises from the conflict between justice and the value of human life: should AVs prioritize law-abiding riders wearing a helmet or those without a helmet whose life is more at risk (Lin 2016)? It is reasonable to want AVs to follow previously agreed-upon ethical principles when making such decisions.

If accidents and AVs are inseparable and high levels of automation are completely human-independent, ethical settings applied to AVs may help create machines that can be a trusted part of the traffic community. Trust in AVs has already been asserted as an essential determinant of the public's receptibility to AVs (Shahrdar et al. 2019; Sharma et al. 2021). The main factors contributing to trust between AVs and their stakeholders identified in the literature are respect for privacy in the data collection necessary for machine learning, cybersecurity, and clear accountability

[1] In January 2023, Mercedes became the first company to achieve certification from SAE for a level 3 automation car (Tarantola 2023).

[2] Such sensors include cameras, stereovision, infrared, radar, ultrasonic, and light detection and radar (LIDAR). The latter is particularly efficient for object detection and distant estimation but has only 90% accuracy in object recognition and suffers from performance decline in adverse weather conditions (Cunneen et al. 2020, 64–65).

[3] Other highlighted benefits of AVs include reducing driving stress and preserving the environment (Dubljević et al. 2021).

[4] Statistics from the National Highway Traffic Safety Administration corroborate this statement, indicating nearly 400 crashes involving level 2 systems happened on US roads between July 2021 and December 2022 (Standing General Order on Crash Reporting For incidents involving ADS and Level 2 ADAS 2021).

in the case of damage (Cunneen et al. 2020; Martinho et al. 2021). Discussed and deliberated ethical settings for AVs should greatly favor the last factor and, consequently, AVs' trustworthiness.

A common objection to the development of ethical AVs concerns their technical infeasibility (Cunneen et al. 2020). To the extent that AV software architecture is inadequate at distinguishing people and other vehicles, these systems are not currently capable of collecting sufficient information to make moral decisions. However, it is reasonable to predict that the current sensors will be implemented with facial recognition, vehicle-to-vehicle (V2V), and vehicle-to-infrastructure (V2I) communications in future years (Cunneen et al. 2020). Such technologies theoretically permit the vehicle to make the fine-grained classifications necessary for ethical decision-making. Another possible concern is the fact that in the case of AV malfunction, hypothetical ethical settings would stop working and, since malfunction is the most common cause of AV accidents (Standing General Order on Crash Reporting For incidents involving ADS and Level 2 ADAS 2021), the ethical settings would be powerless. Nevertheless, the mere possibility of collisions in which ethical settings work and are required is alone a sufficient reason to discuss a plausible ethical framework.

Given the importance of ethical AVs, we contend that understanding how lay people reason ethically about AVs is a foundational task before deliberating the ethical settings. This paper is not the place to fully defend such a methodological claim,[5] but it is worth mentioning that, to design trustworthy AVs, legislators and engineers should avoid applying ethical settings that do not fit with potential stakeholders' moral judgments. Suppose certain moral principles or theories were decided only by a restricted group of experts and applied top-down to AVs' ethical settings. In that case, private consumers may simply "opt out of using AVs thus nullifying all their expected benefits" (Bonnefon et al. 2020, 110). For this reason, we assume here that lay people's intuitions should contribute to informing AVs' ethical settings. For this purpose, it is vital to identify the most solid, stable, and cross-cultural moral intuitions about potential decisions that AVs might make in traffic. To do that, we need a suitable experimental setting that, at the same time, can collect a large number of moral judgments across a wide range of participants.

3 The moral machine experiment and the trolley paradigm

In line with the idea that public morality matters for ethically informed AVs, some scholars have made the moral machine experiment (MME) publicly available (Awad et al. 2018). The goal of this online study was to collect people's moral preferences in traffic scenarios across the world. At the time of publication, it had amassed 39.61 million decisions. To each participant, the experiment displays 13 variations of nine factors, some being the number of people, age, gender, social status, and physicality. Respectively, the five most decisive factors reported were human beings (vs. animals), the number of lives spared, age, compliance with the law, social status, and physical fitness. The breadth of the countries with over 100 participants also enabled the group to create three cultural clusters (Eastern, Southern, and Western) to add a comparative evaluation of the data.

The MME is a valuable starting point for studying moral judgment in traffic because the design is simple and repeatable, making it conducive for gathering large quantities of data and grasping people's preferences. However, the study has also been highly criticized for its naïve assumptions about morality[6] and the unrealistic nature of the proposed dilemmas (Harris 2020; Etienne 2022; Geisslinger et al. 2021). Our criticisms focus on this latter point. Specifically, we contend that high-stakes sacrificial dilemmas employed in the experiment are unfit for investigating moral judgment in traffic scenarios.

The dilemma framework employed in the MME was inspired by the *trolley case*, a famous thought experiment created by Foot (1967) and notably discussed by Thomson (1985).[7] The salient feature of trolley cases is that an agent has to decide to violate a moral norm (e.g., killing one person) to avoid a larger amount of unavoidable harm (e.g., the death of many persons).[8] Such an idealized situation has been utilized for different purposes (Himmelreich 2018, 671–672): as a problem for moral theory (Königs 2022), as a didactical tool, as a social dilemma in the context of AVs (Bonnefon et al. 2016), and as an experimental paradigm for

[5] For more detailed discussions, see Savulescu et al. (2021) and Dubljević (2020).

[6] Specifically, the experimenters illegitimately take personal preferences as moral judgments (Harris 2020).

[7] The trolley dilemma has many variants. In the original version, a bystander has to decide whether diverting a trolley toward a worker on a track to save five workers on the adjacent track. In the famous "footbridge" version, the bystander has to push a fat man off a footbridge to stop a trolley headed toward five men (Thomson 1985, 1409).

[8] We use the expressions "trolley case" or "trolley dilemma" to refer to a type of moral scenario described by such fundamental features. The "trolley problem" is instead the conceptual problem for moral theory to identify the principle that underlies our moral responses to different variants of trolley cases (Königs 2022, 3).

moral judgment. We will not discuss here the importance of trolley cases for ethical theory, its usefulness as a didactical tool, or its relevance as a social dilemma. Instead, our main concern regards trolley dilemmas as an experimental paradigm for moral judgment, that is, as a method to systematically elicit certain kinds of intuition. In particular, moral psychology has extensively employed trolley-like cases to pit consequentialist and deontological intuitions against each other (Greene 2013). In recent years, the trolley dilemma has also emerged as an influential paradigm to study how humans make moral judgments in the cases of AV unavoidable collision, and, arguably, the MME has greatly contributed to enhancing such a paradigm.[9] Despite its merits, we argue that the trolley case is an inadequate experimental paradigm for informing AVs' ethical settings because it is reductively *binary* and lacks *ecological validity*.[10]

Binary choice models are well-suited for experimentation since they enable the cut-and-dry variation of a variable. However, trolley-like dilemmas only permit deontic or utilitarian evaluations, failing to consider other important factors influencing moral judgment. Two relevant unincorporated determinants of moral judgment stemming from virtue ethics are the intertwined concepts of character and intention. Indeed, the importance of character evaluation in moral judgment has been largely documented (Uhlmann et al. 2015). Specifically, some evidence suggests that people may interpret certain stimuli as deeply informative of moral character (e.g., cruelty toward animals as a sign of lack of empathy), even if the action involves no norm violation or harm (Uhlmann et al. 2015, 75–76).

Although one may doubt that AVs have genuine moral agency (Etzioni and Etzioni 2017), there are still good reasons to include characters and intentions in traffic scenarios. First, teaching AVs to identify virtues and vices in human drivers might be useful in the probable scenario in which autonomous and non-autonomous vehicles coexist. For instance, it might be necessary for an AV to distinguish between a negligent-distracted driver, who needs to be avoided, and a malicious driver, who intends to harm other people and needs to be blocked. Second, moral agents might ascribe virtues or vices to self-driving cars according to their *driving style*, which can influence their trust in AVs. This hypothesis is corroborated by some recent studies showing that people tend to attribute moral traits (e.g., dishonesty or cowardice) to artificial intelligence systems, albeit to a lesser extent than to human beings (Gamez et al. 2020). More relevantly to AVs, a preliminary study using a virtual simulator has reported that people's trust in self-driving cars covaries with the AV's driving style (aggressive vs. prudent and respectful) (Shahrdar et al. 2019). Regardless of whether the traits ascriptions based on driving style are genuine virtues and vices, or only their functional equivalents, they still constitute an essential element in traffic moral judgment that undermines the dichotomy between traffic rules and consequences.

As mentioned, the MME's dilemmas include many characteristics irrelevant to agents' moral character. Some of them, such as gender, social status, and physical shape, are probably discriminative and activate implicit biases in the participants. If a substantial amount of collected judgments tend to be biased and discriminative, the overall data do not help inform AVs' ethical settings. Rather, such a goal requires character evaluations based on cues about the moral context embedded in the traffic environment.

Another critical limitation of the trolley paradigm (and the MME particularly) is its lack of *ecological validity*, namely the extent to which some experimental results can be generalized to explain a wide range of real-life situations. In particular, as highlighted (Bauman et al. 2014), experiments based on trolley cases do not have sufficient *experimental*, *mundane*, and *psychological* realism.

Experimental realism concerns how well a study engages participants and makes them take the tasks seriously (Bauman et al. 2014, 537). Trolley-like cases tend to be deficient in this respect because some people find the task humorous rather than serious. This response is probably due to the fact that it is hard to imagine how such a situation can happen in real life. The implausibility of trolley scenarios also negatively affects the psychological realism of the study, that is, the extent to which the study activates in the participants the mental processes involved in real-life moral judgment. In particular, given the humorous effect, trolley-like cases fail to elicit the appropriate moral emotions before imminent death. Likely, this lack of psychological realism affects the MME too: the "video game effect" we feel while choosing whether to kill an old woman or an infant prevents us from having the correct moral attitude that the dramatic nature of the decision would require. Finally, trolley-like cases clearly lack mundane realism because they depict situations considerably far from moral situations than people can find in everyday life. As we will argue in the next section, using plausible mundane traffic scenarios is crucial to inform AVs, and the MME failed to accomplish this objective by creating only high-stakes scenarios. Therefore, the intertwined lack of experimental, psychological, and mundane realism seriously impairs the ecological validity of the trolley paradigm and, hence, its suitability for informing AVs' ethical settings.

[9] According to a recent review (Martinho et al. 2021, 560), more than half of 238 examined articles on AVs ethics mention the trolley case.

[10] Other already highlighted reasons against the use of the trolley paradigm in AVs ethics are the "prospective" decision-making in AVs ethics, the presence of legal responsibility, and the uncertainty of the outcomes in AVs collisions (Nyholm and Smids 2016, Himmelreich 2018).

To summarize, although the MME had the merit to collect a large quantity of worldwide data, it relies on an experimental paradigm that fails to incorporate moral character considerations and suffers from a lack of ecological validity. Therefore, a new experimental paradigm for traffic moral judgments is needed to solve these issues and better inform ethical AVs.

4 Toward more realism: virtual reality and the challenge of mundane traffic situations

For years, studies on moral judgment tested participants by presenting them with abstract textual descriptions of dilemmas. However, with the development of virtual reality (VR) technology, researchers can now employ immersive and naturalistic depictions of morally salient situations to test moral judgment. Arguably, VR scenarios are tremendously beneficial for the ecological validity of a study for different reasons. First, thanks to the vividness and likelihood of the experienced scenarios, it is more probable that participants take moral actions seriously and activate those emotional processes typically elicited by moral situations in real life (Patil et al. 2014). Second, VR, unlike textual vignettes, is less prone to context loss and cognitive biases during language translation, increasing the comparability between moral judgments from different cultures. Third, VR simulators can facilitate a perspective shift from the subjects as mere observers of moral situations to its protagonist. Thus, VR studies can provide more realistic data about decision-making. This aspect is particularly important in the moral domain, given the reported discrepancy between moral judgment and actual behavior (Patil et al. 2014; Rovira et al. 2009).

For all these reasons, a growing number of studies have made use of VR to test moral judgment in road traffic scenarios. Sütfeld et al. (2017) studied people's moral decisions in traffic by observing them driving in a VR simulator. Specifically, in their experimental design, as participants drive a virtual car on a two-lane road, different obstacles appear simultaneously in both lanes: objects, animals, or human beings; the driver has to decide which obstacle to hit. In line with the results of the MME, it turned out that people tend to prioritize sentient beings over objects, humans over animals, many people over one person, and younger people over older ones. Such a consequentialist trend was confirmed by other recent studies using a similar VR setting (Faulhaber et al. 2019; Li et al. 2019), although the introduction of legal elements (e.g., sidewalks and crosswalks) attenuates the prioritization of human life. In contrast with the consequentialist tendency, a preliminary VR study conducted by Grasso et al. (2020) reports that people tend to hit a child crossing the road illegally, instead striking three pedestrians on a sidewalk or two workers in the opposite lane. Interestingly, the distribution of choices drastically shifts toward protecting the child when the virtual driver is not under time pressure and can coldly evaluate her options.

While the aforementioned studies constitute a great advance in ecological validity compared to the MME, they all focus on high-stakes sacrificial dilemmas in which the protagonist has to decide between two sets of lives. As argued, these trolley-like situations suffer from a lack of mundane realism because they are distant from common traffic mishaps. Therefore, although the use of VR helps further mitigate the humorous effect, thus increasing experimental and psychological realism, a more radical departure from the trolley paradigm toward low-stakes traffic scenarios is necessary to increase the mundane realism.

Other scholars (Himmelreich 2018; Borenstein et al. 2019) have already highlighted the importance of mundane traffic situations for the ethics of AVs. One reason for investigating moral decision-making in mundane scenarios is the *challenge of specificity* (Himmelreich 2018, 678): AV algorithms need to know how to act in specific situations according to contextual parameters with a sense similar to human intuitive decision-making. Moreover, studying mundane situations is necessary for tackling the *challenge of scale* (Himmelreich 2018, 678). Since driving behavior in AVs is determined by general policies, decisions made in frequent mundane situations constitute a large-scale problem that needs to be carefully addressed. Last but not least, studying mundane scenarios is necessary for tackling what we call the *challenge of flow*. Since dramatic situations in traffic do not arise without mundane (bad decisions) happening first, AV algorithms trained only on high-stakes traffic scenarios wouldn't be able to detect low-stakes warning signs (e.g., a vehicle frequently changing lanes on a highway without signaling) nor adapt for the possibility of a major event (e.g., a vehicle pile-up).

Another relevant limitation of the mentioned studies is the absence of character and intentions considerations in the employed moral scenarios. The main focus of the VR studies published thus far is to evaluate whether people apply consequentialist reasoning in traffic decision-making. Some of them include traffic norms to observe whether deontological considerations mitigate people's consequentialism. Nevertheless, no study incorporates considerations stemming from virtue ethics, such as if the agent is benevolent, malicious, or negligent. Whether these factors contribute to the overall moral judgment in traffic remains a largely unexplored question.

In sum, VR task environments certainly increase the ecological validity of traffic moral judgments by bringing more experimental and psychological realism. However, the VR studies published thus far remain attached to the trolley paradigm, thus failing to investigate judgments during morally charged mundane traffic situations. To create more informative studies, researchers need to utilize VR and incorporate low-stakes moral problems. Furthermore, no study published to date includes character and intention-based considerations into moral scenarios.

5 A better alternative to the trolley paradigm: the ADC model and its application to AVs

In this section, we consider how the agent-deed-consequences (ADC*)* model of moral judgment can benefit the ethics of AVs. By elaborating upon some points already outlined in Dubljević (2020), we show how the ADC model can be applied to traffic situations and solve the binary of the trolley problem paradigm.

The core tenet of the ADC model is that moral judgment depends on positive or negative evaluations of three different components: the character of a person (the Agent-component, A), her actions (the Deed-component, D), and the consequences brought about in a given situation (the Consequences-component, C) (Dubljević et al. 2018). As some recent evidence suggests (Sattler et al. 2023), the three components are heuristic cues whose processing occurs automatically and simultaneously in one's mind. The overall moral acceptability judgment resulting from such heuristic processing will be positive or negative according to the components' valence. Therefore, the model predicts moral judgments to be positive if all three A-, D-, and C-components are deemed good and negative if all three components are viewed as bad. For example, if a courageous fireman (A+) jumps into a burning house to save an elderly woman (D+) and everyone survives and is healthy and happy (C+), the moral judgment of the situation will be positive (MJ+). Conversely, if a drug dealer (A−) attacks a child (D−) and the child dies (C−), the moral judgment will be clearly negative (MJ−).

It is still unclear how the three components interact when they do not align. It has been suggested (Dubljević et al. 2018; Sattler et al. 2023) that if the character and intentions of an agent are good, and the deed is good, subjects tend to accept or excuse negative consequences because they see them as accidental. For instance, if a courageous fireman (A+) jumped into a burning house to try to save an elderly woman (D+), but she still dies (C−), impartial observers are still likely to praise the agent and the deed, regardless of the consequences. Similarly, if a courageous fireman (A+) attacks (D−) the drug dealer who is trying to harm a child and succeeds in saving the child's life (C+), impartial observers would likely excuse the norm violation, leading to a positive evaluation of the situation (MJ+) (cf. Sattler et al. 2023).

To the extent that the predictions of the ADC model have been confirmed in different moral domains (Dubljević et al. 2018; Sattler et al. 2023), it is reasonable to expect that traffic moral judgments follow similar trends. Specifically, we hypothesize that people evaluate traffic situations according to heuristic cues about the intentions and character of traffic agents (A), compliance or violation of traffic rules (D), and positive or negative outcomes of a traffic action (C). However, applying the ADC model in AVs' ethics requires some preliminary discussion.

The most challenging aspect of incorporating the ADC model in road traffic situations concerns displaying intentions and character in driving performance. We have already highlighted that there are plausible reasons to teach self-driving cars to recognize intentions and virtues. Nevertheless, the first challenge to this task is describing how character evaluations can be made about drivers while at the same time avoiding social biases and intrusions into drivers' personal lives. Arguably, AVstakeholders may not want AV algorithms to judge them according to their morality outside of the traffic context, nor by virtue of some irrelevant social or cultural characteristics (e.g., their occupation, gender, or physical features). Rather, traffic agents deserve to be evaluated according to their character *as traffic agents*, that is, as drivers, passengers, or pedestrians.

Assuming that, the consequent challenge is to explain how an AV can access character-relevant information without infringing other agents' privacy. It stands to reason that character evaluations must be based on fast and subtle physical cues like those available to human drivers. Arguably, at least four sources of information could help an AV evaluate a driver's character and intention. Note that we are not suggesting here that some particular AV policy is necessarily desirable but only considering different sources of information across various possible AV scenarios for evaluating character on the road.

First, thanks to V2V communication, an AV may access the characteristics of the vehicle it is facing. If the identified vehicle is, for example, an ambulance or a firetruck, the machine can interpret its behavior as benevolent; by contrast, if the vehicle is reported as being stolen, the driver could be considered malicious. Second, in a mixed scenario in which AVs coexist with human drivers, self-driving cars may access the incident history of the human drivers to assess their recklessness. Third, in the scenario that AV owners can choose their ethical settings (e.g., altruistic vs. egoistic),[11] self-driving cars can (and perhaps should) communicate

[11] This normative scenario has been proposed by (Contissa et al. 2017).

Table 1 A tentative framework of virtues and vices of drivers according to their behavior and the source of information from which an AV can judge

Virtue/vice	Behavior	Source
Benevolent/malicious	The driver tends to help other agents on the road/ harm other agents	Vehicle type, driving style
Responsible/irresponsible	The driver takes many risks/ does not take any excessive risk	Driving style, vehicle condition, incidents history
Careful/negligent	The agent drives straightforwardly and reliably/ is distracted and makes zigzags	Driving style, vehicle condition, incidents history
Kind/rude	The driver allows other cars to merge, pedestrians to cross/ does not allow to merge or cross, hurries other drivers	Driving style

A similar framework could be outlined for pedestrians or other traffic agents

their ethical character to each other. Finally, and aligning most with privacy, the driving style exhibited on the road might constitute a relevant cue to determine the virtues and vices of a driver as a driver. For example, a car weaving in and out of traffic can be a sign of a negligent driver; by contrast, a car driving straightforwardly and reliably might indicate that the driver is careful (Table 1).[12]

Of course, driving style should be considered as limited and defeasible evidence for character and intentions on the road. For example, a car driving aggressively, weaving through traffic and sounding the horn to reach the emergency department of the hospital because of a medical crisis could be perceived as vicious despite the good intent.[13] However, like the way hazard lights are used to display a positive intent in such an emergency scenario, it would be ideal for vehicles to communicate their intent to other traffic agents via V2V communication. Using transmittable signals to display distress or urgency is a viable option, but one that comes at the cost of possible user abuse and regulatory hurdles. On the other hand, if given the ability to read patterns of blinking lights (driving style), like its human counterpart, intent can be inferred physically. Admittedly this solution is less feasible and more error-prone. Either way character display recognition is a technological possibility that would certainly enhance AV decision making.

In line with the trolley case trend over the last half-century, operationalizing the D− and C− components is much easier than incorporating traffic agency. An action may count as a positive deed when it complies with a traffic norm and a negative when it does not.[14] Consequences are positive or negative depending on whether any deed results in a collision (C−) or the vehicle reaches its destination without any accident during the route (C+).[15]

Having defined what the A- D- C-components mean in the context of road traffic and identified possible sources of information that represent them, we can now outline the main empirical hypotheses of the ADC model for traffic moral judgments. In line with the results of previous studies on other moral domains (Dubljević et al. 2018; Sattler et al. 2023; Dubljević and Racine 2014), we predict that every single component, if positive, increases the overall moral acceptability of a traffic situation. Therefore:

1. Positive agency results in more positive moral judgments as compared to negative agency.
2. Positive deeds result in more positive moral judgments as compared to negative deeds.
3. Positive consequences result in more positive moral judgments as compared to negative consequences.

Although each component is effective (either positively or negatively), it is likely that one factor has a major influence and determines the valence of the overall judgment. In line with previous results (Dubljević et al. 2018; Sattler et al. 2023), since the scenarios are low-stake, we expect that agency (A) and deeds (D) outweigh the importance of the consequences (C).

6 A proposal for experimental design

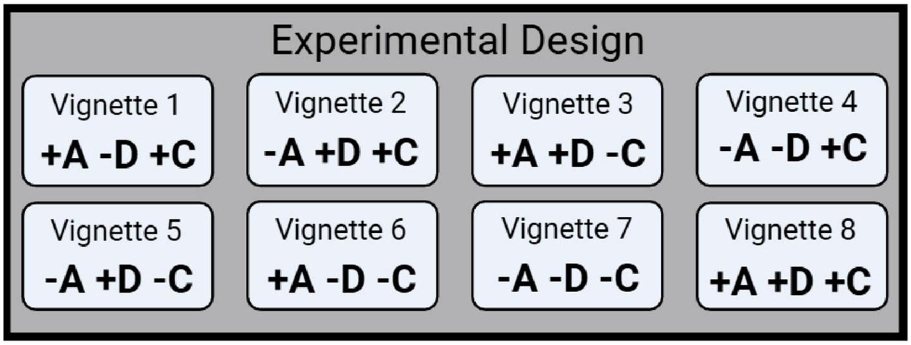
Here we propose an experimental setting to test our hypotheses on traffic moral judgments outlined in the previous section. This original design depicts mundane low-stakes traffic situations operationalized according to the ADC model using VR technologies. We argue that all such elements merged together should increase future studies' ecological

[12] As suggested by Dubljević (2020, 2468), the "Identification Friend or Foe" technology used in military engagements could help an AV to interpret good or bad intentions on the road.

[13] We are grateful to an anonymous reviewer for providing this example.

[14] The deed can be low-stakes (e.g., running a stop sign on an empty rural road) or high-stakes (e.g., driving in the opposite direction on a highway), depending on the severity of the rules infraction.

[15] The consequences can also be high-stakes if the life of a person is at risk or low-stakes in the case of mundane situations.



Experimental Design

Vignette 1 +A -D +C

Vignette 2 -A +D +C

Vignette 3 +A +D -C

Vignette 4 -A -D +C

Vignette 5 -A +D -C

Vignette 6 +A -D -C

Vignette 7 -A -D -C

Vignette 8 +A +D +C

Fig. 1 (Created in BioRender) The moral aspect combinations for each vignette within the larger scenarios

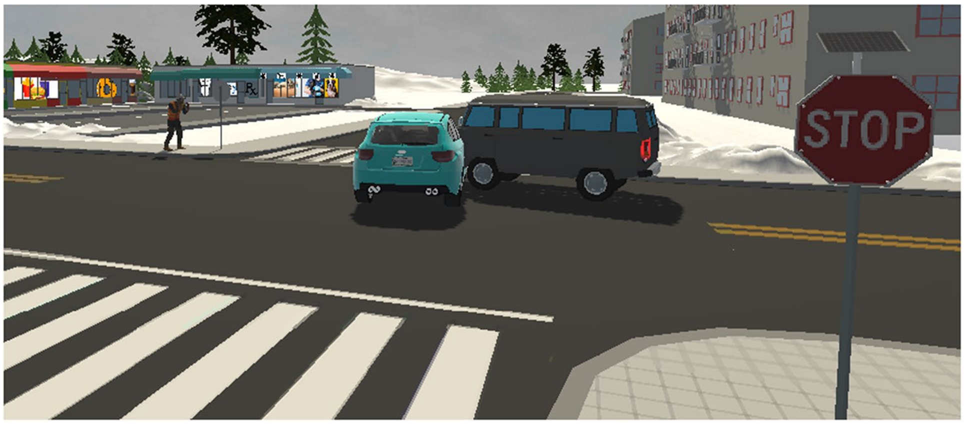


Fig. 2 The frame in which the protagonist car hits a truck running across the orthogonal road (D−, C−)

validity, helping the moral psychology of AVs move beyond the trolley paradigm.

6.1 Methodology and procedure

We developed six low-stakes virtual traffic scenarios. Each scenario gives rise to eight different vignettes depending on the combination of the three (A-D-C) components and their valence (Fig. 1). All the variations in the scenarios are composed of unique situations deemed morally salient and contextually clear representations of the ADC model (see also Dubljević et al. 2018).

The structure of each scenario is as follows: a traffic agent displays some form of virtue or vice (A+/A−), then obeys or disobeys some traffic rule (D+/D−) that finally results in some positive or negative consequence (C+/C−). For example, in one of our developed scenarios, the protagonist car is approaching an intersection regulated by a stop sign. The driver either tends to help other traffic agents (A+) or tends to harm other traffic agents (A−). Then, the story differs according to the combination of good or bad deeds and positive or negative consequences:

> (D+, C+) The car correctly stops before the sign avoiding a collision with a truck running across the orthogonal road, and the protagonist arrives on time at work;
>
> (D−, C+) The car cannot stop before the sign for some reason (in the A− version because of the driver's negligence and in the A+ version because of an animal suddenly forcing the main car to swerve), but no collision occurs and the protagonist arrives on time at work;
>
> (D+, C−) The car correctly stops before the sign, but a collision occurs because a truck coming from the opposite side swerves into the wrong lane, hitting the protagonist car;
>
> (D−, C−) The car cannot stop before the sign for some reason (depending on the valence of the A) and collides with a truck running across the orthogonal road (Fig. 2).

The six scenarios were created using the Unity Real-Time Development Platform version 2020.3.18f1 and are deployable on three VR technologies. These technologies include low-immersion desktop videos for large-scale data collection through Amazon Mechanical Turk, 360-degree room-scale versions designed for an immersive Visualization Gallery, enabling medium-scale high-resolution surveying (20 participants at a time), and high-immersion Oculus Quest 2 head-mounted display variants for individual assessments. The virtual stimuli, comprising the environment assets, characters, and animations, are files in the public domain downloaded from open-source websites like the Unity-Asset store and Mixamo.com or those created by our laboratory developers using Unity and Blender.

To help ensure that the scenarios we intend to use in the data collection phase of this experiment are contextually clear and accurate depictions of the ADC moral aspects, we are performing an extensive pilot testing phase. Over a year-long period, one hundred and twenty-five subjects were shown half of the scenarios from all six vignettes using head-mounted displays or the Visualization Gallery mentioned previously, totaling eighteen scenarios per respondent. The participant composition included twelve high school students, eighty undergraduate students, seventeen graduate students (one with an extensive emergency response background), and sixteen professional academics. After each piloting session, participant feedback, deemed valid by our research team, was incorporated into the lacking scenarios in an iterative improvement process.

The large-scale data collection will be performed using the desktop video versions and Amazon Mechanical Turk.[16] This shift from more immersive technologies for pilot testing to a less immersive desktop setting for data collection warrants justification. Although desktop versions are less immersive than head-mounted displays, evidence indicates that both elicit high levels of psychological presence (Shu et al. 2019). Additionally, the desktop investigation has practical benefits above the other two technologies. First, desktops produce far less simulation sickness, and second, the infrastructure is already in place to carry out global sampling while still creating greater psychological realism than the crude pictorial representation used in the MME or other textual methodologies.

After they observe a scenario, the participants will be asked to provide four moral acceptability ratings (from 1 to 10) for the version they watch: an overall moral acceptability judgment and three other evaluations corresponding to the three model aspects (“Can the protagonist be described as bad?”, “Can the action be described as bad?” and “Can the outcome be described as bad?”, respectively). To avoid contrast, anchoring, and update effects, we adopt a *between-subject* design in which each participant evaluates only one randomly assigned variant of the eight versions of traffic scenarios (see Fig. 1).

6.2 Discussion

The limitations of the MME and the trolley paradigm collectively point to the need for a revision in experimental design for traffic moral judgment. Our experimental adjustments constitute a great improvement over the trolley paradigm in several respects.

Thanks to the inclusion of temporality and audio-visual stimuli a VR task environment provides, our study’s ecological validity is much greater than that of the MME. Furthermore, the focus on mundane traffic vignettes that frequently happen in real life increases the overall realism and credibility of the study. It is worth noting that it is still possible to set more dramatic moral options through the ADC model by simply increasing the stakes (from low to high) of all components. Nevertheless, given the underestimated importance of mundane situations, we left high-stakes scenarios for future studies.

Another significant improvement brought by the proposed framework is the integration of the character point of view into the context of a traffic situation. Importantly, the ADC model incorporates the three main ethical pillars of moral philosophy: virtue ethics, deontological ethics, and utilitarianism, operationalized into relevant moral stimuli. This means that our experimental design, unlike studies based on the trolley paradigm, can include considerations of character contributing to moral judgment besides norms and consequences-based considerations. Character heuristic cues are based on driving style (reliable driving vs. weaving, or tendency to help vs. harm other traffic agents) and thus are well integrated with traffic action and not dependent on social prejudices.

Our proposed experimental design also has a simple and repeatable structure, apt to be applied for large-scale global investigation. This means that our setup, like the MME, can collect large quantities of data and obtain solid moral preferences necessary to avoid a methodological regression. Our data should also be comparable across different cultures, seeing as the developed virtual scenarios involve minimal language, apart from traffic signs, short written phrases, and small segments of voice acting that can be easily swapped between languages.[17]

Despite the highlighted features, our experimental proposal has some identified limitations. The most significant one is that the subjects participating in the experiment are only observers and not agents. As a result, the collected preferences are mere moral judgments and not decisions like the ones reported in studies using a virtual drive simulator (Sütfeld et al. 2017; Faulhaber et al. 2019; Li et al. 2019; Grasso et al. 2020). Therefore, the goal of future studies will be to test whether people’s decisions in traffic respond to agency, deed, and consequences heuristic cues.

[16] Amazon Mechanical Turk is a website that facilitates payment for completing surveys, and such samples have been shown to provide more reliable and representative data compared to student samples (Buhrmester et al. 2011).

[17] An example of one of the contextual phrases provided is “Remy, I am moving out of my house tomorrow morning. Do you mind coming over at 7:30 with your truck to help me out?”.

Another limitation concerns the order of the moral stimuli, which is predetermined and fixed: first agency, then the deed, and eventually the consequences. A vignette in which a subject obtains such information in a semi-randomized order or almost simultaneously is closer to real life. This predictability in our design enables consistency, but ethical AVs will not have this luxury, so future experiments could look to find a way to manipulate component order.

Moreover, A- D- C-components carry different specific weights in real moral situations. Thus, the acceptability rating of the overall moral judgment we are probing likely depends heavily on the weight of the components. For example, a malicious driver who intentionally harms pedestrians is a much stronger A− than a negligent driver; ignoring a red traffic light could be perceived as a higher D− than not respecting a stop sign. Perhaps, assigning a specific weight to the three components could help make more fine-grained predictions about the interplay between components. While we are partially mitigating this concern by collecting itemized moral ratings for each aspect, we have no specific hypotheses about the aspects' weights. Our current experimental design assumes that each component has a similar weight to avoid excessive complications at this early research stage.

In our proposed scenarios, a subject can clearly observe the actual consequences of the traffic action. Our experimental setting would be closer to reality if the consequences were uncertain and the subjects made judgments based on *expected consequences* or *risk*, which are arguably an important factor in the ethics of AVs (Geisslinger et al. 2021).

Finally, another identified limitation is that our study proposal does not ask participants to express a preference between different AV policies; for example, between a scenario in which AVs coexist with human drivers and one in which only AVs are allowed. In other words, our study proposal focuses only on individual behaviors and not on system-level judgments (Borenstein et al. 2019; Dubljević et al. 2021).

7 Conclusion

AVs will probably bring more safety on the road by significantly reducing human error. However, given the traffic unpredictability, it is impossible to expect that AVs will not be involved in collisions. Thus, self-driving cars will have to learn to make moral decisions before the public can trust them. Presupposing the importance of human moral psychology to inform AVs' ethical settings, the MME attempted to understand people's moral preferences by relying on trolley case-like scenarios. Nevertheless, the experimental design involved is unrealistically binary and lacks ecological validity. Other studies, while improving experimental realism, still rely on the trolley paradigm, failing to incorporate character considerations and mundane moral problems.

To address the highlighted issues, we proposed an original experimental design based on the application of the ADC model of moral judgment. The developed traffic scenarios comprise heuristic cues about the driver's character, her compliance or violation of traffic rules, and the consequences brought by her action. Compared to the MME, the ecological validity of the study has advanced thanks to the use of a VR environment. Furthermore, the focus on low-stakes moral scenarios fills an important gap in the literature on traffic judgments. All these elements combined together create a simple and repeatable structure suitable for collecting a large quantity of data.

More work needs to be done to better understand human moral psychology in road traffic situations. Specifically, future studies should test whether A- D- C-components guide people's decisions in traffic and not just their judgment as observers. Further complications worth examining in future experiments concern the components' weight and the outcomes' uncertainty.

Acknowledgements Earlier versions of this paper were presented at the 2023 *Meeting of the North Carolina Philosophical Society,* the 2023 *Meeting of the South Carolina Society for Philosophy*, and the 9th *Conference of the Italian Society for Neuroethics*. We are extremely thankful to the audience of these events for their precious feedback. We are also grateful to our colleagues from the *NeuroComputational Research Group* at NC State and an anonymous referee of this journal for their suggestions.

Funding Work on this article was supported by a CAREER grant from the National Science Foundation (Awarded to the corresponding author, Grant Number: 2043612).

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest No conflict of interests.

References

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

Bauman WC, McGraw PA, Bertels MD (2014) Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology. Soc Pers Psychol Compass 8(9):536–554

Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6397):36–37

Bonnefon J-F, Shariff A, Rahwan I (2020) The moral psychology of ai and the ethical opt-out problem. In: Liao SM (ed) Ethics of artificial intelligence. Oxford University Press, Oxford, pp 109–126

Borenstein J, Herkert JR, Miller KW (2019) Self-driving cars and engineering ethics: the need for a system level. Sci Eng Ethics 25:383–398

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6(1):3–5

Contissa G, Lagioia F, Sartor G (2017) The ethical knob: ethically-customisable automated vehicles and the law. Artif Intell Law 25(3):365–378

Cunneen M, Mullins M, Murphy F, Shannon D, Furxhi I, Ryan C (2020) Autonomous vehicles and avoiding the trolley (dilemma): vehicle perception, classification, and the challenges of framing decision ethics. Cybern Syst 51(1):59–80

Dubljević V (2020) Toward implementing the ADC model of moral judgment in autonomous vehicles. Sci Eng Ethics 26:2461–2472

Dubljević V, Racine E (2014) The ADC of moral judgment: opening the black box of moral intuitions with heuristics about agents, deeds, and consequences. AJOB Neurosci 5(4):3–20

Dubljević V, Sattler S, Racine E (2018) Deciphering moral intuition: how agents, deeds, and consequences influence moral judgment. PLoS ONE 13(10):e0206750

Dubljević V, List G, Milojevich J, Ajmeri N, Bauer WA, Singh MP, Bardaka E et al (2021) Toward a rational and ethical sociotechnical system of autonomous vehicles: a novel application of multi-criteria decision analysis. PLoS ONE 16(8):e0256224

Etienne H (2022) A practical role-based approach for autonomous vehicle moral dilemmas. Big Data Soc 1–12

Etzioni A, Etzioni O (2017) Incorporating ethics into artificial intelligence. J Ethics 21:403–418

Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, Stephan A, Pipa G (2019) Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. Sci Eng Ethics 25:399–418

Foot P (1967) The problem of abortion and the doctrine of the double effect. Oxf Rev 5:5–15

Gamez P, Shank BD, Arnold C, North M (2020) Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. AI Soc 35:795–809

Geisslinger M, Poszler F, Betz J, Lütge C, Lienkamp M (2021) Autonomous driving ethics: from trolley problem to ethics of risk. Philos Technol 34:1033–1055

Goodall NJ (2014) Ethical decision making during automated vehicle crashes. Transport Res Rec 2424:58–65

Goodall NJ (2018) How to think about driverless vehicles. Am J Public Health 108(9):1112–1113

Grasso GM, Lucifora C, Perconti P, Plebe A (2020) Integrating human acceptable morality in autonomous vehicles. In: Ahram T, Karwowski W, Vergnano A, Leali F, Taiar R (eds) 3rd international conference on intelligent human systems. Springer, Modena, pp 41–45

Greene J (2013) Moral tribes: emotions, reason, and the gap between us and them. The Penguin Press, New York

Harris J (2020) The immoral machine. Camb Q Healthc Ethics 29:71–79

Hawkins AJ (2022) Theverge. https://www.theverge.com/2022/1/5/22867455/gm-sell-autonomous-vehicles-personally-owned-timeline. Accessed 3 Apr 2023

Himmelreich J (2018) Never mind the trolley: the ethics of autonomous vehicles in mundane situations. Eth Theory Moral Pract 21:669–684

Königs P (2022) Of trolleys and self-driving cars: What machine ethicists can and cannot learn from trolleyology. Utilitas 35(1):70–87

Li S, Zhang J, Li P, Yongqing W, Wang Q (2019) Influencing factors of driving decision-making under the moral dilemma. IEEE Access 7:104132–104142

Lin P (2016) Why ethics matters for autonomous cars. In: Maurer M, Gerdes JC, Lenz B, Winner H (eds) Autonomous driving: technical, legal and social aspects. Springer, Heidelberg, pp 69–86

Lütge C, Poszler F, Acosta AJ, Danks D, Gottehrer G, Mihet-Popa L, Naseer A (2021) AI4People: ethical guidelines for the automotive sector—fundamental requirements and practical recommendations. Int J Technoeth 12(1):101–125

MacRumors Staff (2023) MacRumors. https://www.macrumors.com/roundup/apple-car/. Accessed 3 Apr 2023

Martinho A, Herber N, Kroesen M, Chorus C (2021) Ethical issues in focus by the autonomous vehicles industry. Transp Rev 41(5):556–577

Millar J (2017) Ethics settings for autonomous vehicles. In: Lin P, Keith A, Jenkins R (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, New York, pp 20–34

Nyholm S, Smids J (2016) The ethics of accident-algorithms for self-driving cars: an applied trolley prolem? Eth Theory Moral Pract 19:1275–1289

Patil I, Cogoni C, Zangrando N, Chittaro L, Silani G (2014) Affective basis of judgment behavior discrepancy in virtual experiences of moral dilemmas. Soc Neurosci 9(1):94–107

Rovira A, Swapp D, Spanlang B, Slater M (2009) The use of virtual reality in the study of people's responses to violent incidents. Front Behav Neurosci 3:59

Sattler S, Dubljević V, Racine E (2023) Cooperative behavior in the workplace: empirical evidence from the agent-deed-consequences model of moral judgment. Front Psychol 13:1064442

Savulescu J, Gyngell C, Kahane G (2021) Collective reflective equilibrium in practice (CREP) and controversial novel technologies. Bioethics 35:652–663

Shahrdar S, Park C, Nojoumian M (2019) Human trust measurement using an immersive virtual reality autonomous vehicle simulator. In: AIES'19: session 9: human and machine interaction. Association for Computing Machinery, Honolulu, pp 515–520

Sharma O, Sahoo N, Puhan N (2021) Recent advances in motion and behavior planning techniques for software architecture of autonomous vehicles: a state-of-the-art survey. Eng Appl Artif Intell 101:104211

Shu Yu, Huang Y-Z, Shu-Hsuan C, Chen M-Y (2019) Do virtual reality head-mounted displays make a diference? A comparison of presence and self-efcacy between head-mounted displays and desktop computer-facilitated virtual environments. Virtual Real 23:437–446

Singh S (2018) Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. Traffic Safety Facts Crash•Stats. National Highway Traffic Safety Administration, Report No. DOT HS 812 506, Washington, DC

Standing General Order on Crash Reporting For incidents involving ADS and Level 2 ADAS (2021) General Order. National Highway Traffic Safety Administration, Washington, D.C.

Sütfeld LR, Gast R, Koenig 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. Front Behav Neurosci 11:122

Tarantola A (2023) engadget. https://www.engadget.com/mercedes-first-certified-level-3-autonomy-car-company-us-201021118.html

Thomson JJ (1985) The trolley problem. Yale Law J 94(6):1395–1415

Uhlmann LE, Pizarro AD, Diermeier D (2015) A person-centered approach to moral judgment. Perspect Psychol Sci 10(1):72–81

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.