Moving Away from The Trolley Problem

A paradigm shift in ethics of autonomous vehicles

1. Introduction

Low-level autonomous vehicles (AVs), which either assist humans with some aspects of navigation or require continuous supervision, are already being sold in the U.S., and the deployment of higher-level AVs in the future is imminent (Canis 2019). While behavior guidance systems (the hardware and software architectures that enable intelligent motion planning) are rapidly advancing, the rules that underpin these systems are technically and ethically insufficient for the feasible incorporation of higher-level AVs, which are entirely autonomous (Sharma, Sahoo and Puhan 2021). Since these rule sets underscore many ethical issues, including safety and accountability, some moral philosophers and engineers hope to inform AV systems of human drivers' moral intuitions during traffic dilemmas (Lin 2016).

In recent years, the Moral Machine Experiment (MME) has emerged as a very influential paradigm for investigating moral judgment in traffic scenarios (Awad, et al. 2018). Although it had the merit to collect millions of data from subjects across the world, the MME has many limitations (Harris 2020, Cunneen, et al. 2020, Etienne 2022). In particular, the employment of high-stakes sacrificial dilemmas inspired by the trolley problem thought experiment is problematic because they are simplistically binary and lacks external validity.

A paradigm shift in the study of traffic moral judgments is in order to inform future research in AV ethics and eventually to create aligned software. Our original solution comprises the creation of realistic traffic moral scenarios by using virtual reality and the agent-deed-consequence (ADC) model of moral judgment as a moral-psychological framework (Dubljević and Racine 2014, Dubljević, Sattler and Racine 2018, Dubljević 2020).

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 (section 2). Then, we highlight the main flaws of the MME in

section 3. Finally, in section 4, we describe our experimental design and explain how it corrects the faults pertaining to the MME.

2. The Importance of Ethically Acting Autonomous Vehicles

An AV is a vehicle that can fully or partially drive without a human operator for a prolonged time.[1] AV software is ethically informed whenever it responds to environmental stimuli by following previously agreed-upon ethical rules. 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, Sahoo and Puhan 2021). Ethical AVs would be equipped with *Ethics Settings* (ESs), defined as *modes of ethical choice-making embedded in the vehicle's planning layer* (Millar 2017). We think that ESs are necessary for the development of AVs 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 its autonomous decisions.

An estimated 94% of traffic accidents were attributed to human error in 2015 (Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey 2015). Thus, the implementation of AVs will have the potential benefit of reducing the considerable harm caused by car accidents. Other highlighted benefits of AVs include reducing driving stress and preserving the environment (Dubljević, List, et al. 2021). However, accidents with AVs do still happen.[2] Malfunctions will always be possible, and the coexistence of autonomous and non-autonomous cars increases the probability of an accident. Thus, it is likely that AVs will have to choose between two unfavorable alternatives: for example, prioritizing pedestrians instead of passengers or humans

[1] The Society of Automotive Engineers identifies six levels of automation: from level 0 (No automation) to level 5 (full automation). To date, only up to level-2 AVs occupy roadways (Canis 2019).

[2] 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).

instead of animals (Lin 2016). It is reasonable to want AVs to follow plausible ethical principles when making such decisions.

If accidents and AVs are inseparable and the higher levels are to be completely human-independent, ESs may help create machines that can be a trusted part of our traffic community. Trust in AVs has already been asserted as an essential determinant of the public's receptibility to AVs (Sharma, Sahoo and Puhan 2021; Gopinath and Narayanamurthy 2022). As has been suggested (Pflanzer, et al. 2022), people display trustworthiness when they meet three conditions; ability (competence), benevolence (right actions), and integrity (honesty and consistency). When applying this framework to AVs, the ability component is cited as the most important of the three but is primarily a byproduct of engineering; nevertheless, functional ESs could increase the other two factors more drastically.

Given the importance of ethical AVs, we think that understanding how human beings reason ethically about AVs is a crucial task. This is not the place to provide a full defense of such a methodological claim,[3] but it is worth mentioning that to design trustworthy AVs, legislators and engineers should avoid applying ethical rules that do not fit with potential consumers' moral judgments. For this reason, we contend that lay people's intuitions should contribute to informing AV 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 realistic 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 (MME) and its troubles

In line with the idea that public morality matters for ethically informed AVs, some scholars have made publically available the MME (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 36.1 million decisions from across the globe. To each participant, the experiment displays 13 variations of nine factors, some

[3] See Dubljević (2020) for a more detailed discussion.

being: age, gender, social status, and physicality. Respectively, the five strongest preferences reported were humans over animals, the number of lives spared, the young, law-abiders, and those of perceived higher status. 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 useful starting point for studying moral judgment in traffic because its simple and repeatable design is conducive to gathering large quantities of data and grasping people's preferences. However, the study has also been highly criticized for its naïve assumptions about morality and the unrealistic nature of the proposed dilemmas (Harris 2020, Cunneen, et al. 2020, Etienne 2022). Our criticisms focus on this latter point. Specifically, we contend that high-stakes sacrificial dilemmas employed in the experiment (figure 1) are unfit for investigating moral judgment in traffic scenarios.

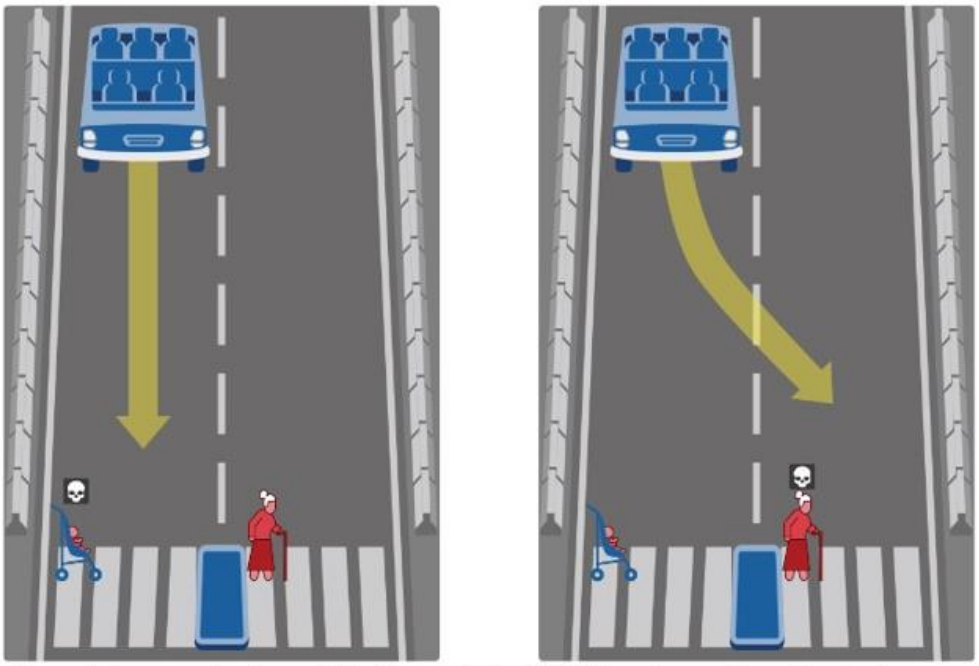

Figure 1. An example of sacrificial dilemma in the MME. Here the participant has to choose whether to sacrifice an old woman or an infant.

The dilemma framework used in the MME was inspired by the *trolley problem*, a prominent thought experiment created by Philippa Foot (1967) and notably discussed by Thomson (1985). The dilemma has many variants, but all of them involve the possibility of an agent saving many lives by killing a person. The trolley problem has been extensively employed in moral psychology to pit consequentialist and deontological

intuitions against each other (Greene 2013). Despite its simplicity, we argue that this type of moral scenario is inadequate for informing ESs because it is simplistically binary and lacks external validity.

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 the field of virtue ethics are the intertwined concepts of character and intention. Indeed, the importance of character evaluation in moral judgment has been largely documented (Uhlmann, Pizarro and Diermeier 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, Pizarro and Diermeier 2015, 75-76). One could object that the perception of character is irrelevant to the ethics of AVs to the extent that AI systems do not have a character like humans do. Nevertheless, recent studies suggest that people attribute moral traits (e.g., dishonesty or cowardice) to AI systems, albeit to a less extent than to human beings (Gamez, et al. 2020). Therefore, excluding the character point of view is an important limitation in the study of the moral judgment of AVs.

As has been highlighted (Bauman, McGraw and Bertels 2014), trolley problem-based experiments also lack substantial *external validity*, namely the extent to which the results of a study can be generalized to explain a wide range of situations. Specifically, the trolley problem variants do not have sufficient *experimental*, *mundane*, and *psychological* realism. Given the humorous effect they elicit, subjects do not tend to take trolley dilemmas seriously, causing them to fall short of experimental realism. Furthermore, these scenarios lack mundane realism because they are far from the moral situations one can face in real life. Finally, studies employing trolley dilemmas are not psychologically realistic: as they fail to activate in the participants the mental processes typically involved in real moral judgments. Likely, this lack of overall 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.

4. A new paradigm in the study of moral judgment in traffic scenarios

The limitations of MME and sacrificial dilemmas collectively point to the need for a revision in experimental design. We propose an original experimental setting for testing people's moral judgment in traffic scenarios. Our experimental design is based on the application of the agent-deed-consequence (ADC) model of moral judgment (Dubljević and Racine 2014, Dubljević, Sattler and Racine 2018, Dubljević 2020) and the use of virtual reality (VR). We argue that this movement away from the trolley problem and toward increased virtual realism removes the binary choice issue and produces higher external validity; while maintaining the ability to collect cross-cultural data and creating a synergy between manageable operationalizability and explanatory power. The remainder of this section describes our experiment and asserts why the ADC model is better than other moral models for guiding the creation of future ESs (4.1) and shows how VR can facilitate empirical research not possible in the past (4.2).

4.1. The ADC Model and Its suitability for guiding the Creation of ESs

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), their actions (the Deed-component, D), and the consequences brought about in a given situation (the Consequences-component, C) (Dubljević and Racine 2014, Dubljević, Sattler and Racine 2018). Recent cross-cultural evidence suggests that the appraisal of these components occurs in one's mind simultaneously and automatically (Sattler, Dubljevic and Racine 2022). The overall moral acceptability judgment resulting from this process will be positive or negative according to the components' valence and weight. For example, the model predicts moral judgments to be positive if all three of the A-, D-, and C-components are deemed good and negative if all situation characteristics are viewed as bad. If the components do not align, the valence of the resultant judgment will depend on the specific weight of the A-, D-, and C-components.

We predict that in line with previous evidence in other moral domains (Dubljević and Racine 2014, Dubljević, Sattler and Racine 2018), people's moral judgments in traffic situations rely on evaluations of A-, D-, and C-components. To test this hypothesis, we have developed six low to medium-stakes virtual traffic dilemmas that pit positive and

negative moral aspects against each other, totaling eight possible logical arrangements (Figure 2). The structure of each vignette is as follows: An agent displays some form of virtue/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 six developed scenarios, a husband (**A**) is either attentive/nurturing or violent/demeaning to his pregnant wife as they get into their car to drive to the hospital for delivery. In the following scene, while driving on a foggy rural road, the agent is forced to negotiate a roadway obstacle in the form of one large stationary herd animal in the agent's lane (Figure 3). In half of the scenarios, the agent follows traffic laws by not crossing a double yellow line, while in the other half, he disregards traffic rules by swerving into the oncoming lane (**D**). At this point, the positive and negative deeds bifurcate into safe navigation of the obstacle or an accident (**C**). The variations at this branch point are presented below:

- (**D+**, **C+**) The agent stops on time in the correct lane avoiding an accident and allowing the couple to make it to the hospital on time for labor.
- (**D-**, **C+**) The agent drives past the cow in the oncoming lane without slowing down, allowing the couple to make it to the hospital on time for labor.
- (**D+**, **C-**) The agent fails to stop, colliding with the cow, resulting in the wife's miscarriage.
- (**D-**, **D-**) The agent moves into the oncoming lane as a vehicle approaches, causing an unshown implied accident, resulting in the wife's miscarriage.

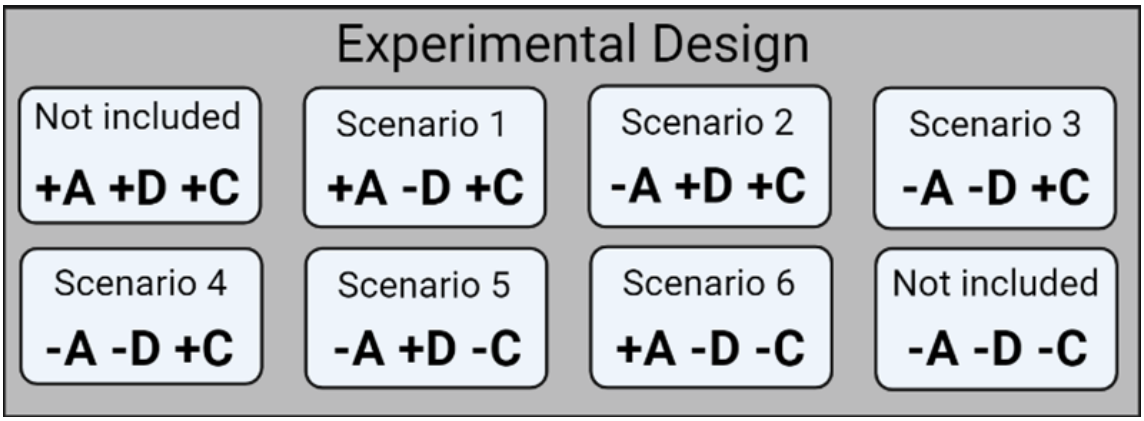

Figure 2. (Created with BioRender) Shows the possible arrangements created when varying the three model components over the positive/negative dimension, presented in the experimental order. The entirely positive and negative versions of the scenario are excluded in line with prior design justifications and to perverse programming and participant resources (Dubljević, Sattler and Racine 2018).

Figure 3. Is a screenshot taking of one of the four deed-consequence versions of our virtual cow vignettes from the perspective of the camera. In this scenario, the agent's vehicle stops in time to avoid colliding with the cow (**D+**, **C+**)

Importantly, the ADC model incorporates the three main ethical pillars of moral philosophy: virtue ethics, deontology, and utilitarianism, operationalized into relevant moral stimuli. That means that our experimental design, unlike the MME, can include considerations of character contributing to moral judgment. Furthermore, despite being non-binary, the framework adopted is still easy to operationalize and thus permits collecting a large amount of comparable data. Finally, another advantage of our experimental framework is that it leaves room for low- and medium-stakes scenarios in which no human life is at risk. These aspects make the dilemmas closer to everyday traffic

situations, increasing the mundane realism of the study relative to the typical sacrificial dilemma.

4.2. Design improvements due to VR Task Environments

For all six vignettes, our team has created three levels of virtual stimuli using the Unity Real-Time Development Platform version 2020.3.18f1. These include low-immersion desktop videos for large-scale data collection through Amazon Mechanical Turk, 360-degree room-scale versions designed for our university's Visualization Studio, enabling medium-scale surveying (20 participants at a time), and high-immersion Oculus Quest 2 variants for individual assessments. The virtual stimuli, comprising environment assets, characters, and animations, are mostly free-to-use files downloaded from the Unity-Asset store or Mixamo.com, with some components like the English voice acting created entirely in-house.

Arguably, the use of VR is highly beneficial to a study's external validity. The immersive nature of a virtual environment increases experimental and psychological realism (Smith 2015). This increase in the feeling of realism is attributed to the psychological notion of presence in the virtual space, enabling a more accurate examination of the phenomenon under study compared to surveys of the same content. Certainly, no participant believes the virtual environment is real, but the 3-dimensional audiovisual aspects of VR will get our experiment closer to testing authentic moral judgment processes. 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.

In general, textual vignettes are prone to context loss during language translation, limiting their suitability for guiding the creation of AV ESs. The schematic pictures employed in the MME are a step toward language independence. However, as argued, they are too simplistic and unrealistic. Our virtual moral scenarios address this limitation and facilitate data acquisition from non-native English speakers.

Figure 4. Illustrates the level of realism displayed in our virtual vignettes. *Above*, depicts a Fire Marshall blocking a crowd of angry individuals, preventing them from chasing a victim in an urban environment. *Below*, exhibits the scenarios' agent being rear-ended after appropriately stopping for a stop sign in a snowy rural environment.

5. Conclusion

Low-level AVs currently occupy our roadways, and following the progression of this adoption trend, higher-level versions will also be integrated into the traffic system. However, given present software and hardware architectures, it is impossible to expect that AVs will not be involved in collisions, both serious and mundane. Thus, the tension between their ability to increase safety and the inevitable harm they will cause forces engineers and manufacturers to answer a moral question before the public can trust these machines. Presupposing the importance of human moral psychology to address this question, the MME attempted to understand people's moral preferences by relying on trolley problem-like scenarios. Nevertheless, the experimental design involved is

unrealistically binary and lacks experimental validity. We addressed these flaws by proposing an original experimental setting to test people's moral judgments in realistic traffic scenarios. Our proposal accommodates character evaluations besides rule- and consequences-based evaluations. Moreover, using VR, the external validity of the results has likely increased. In sum, we have created an experimental design that maintains the trolley problem core while eliminating glaring issues, adding nuance to the results, and broadening the participant pool.

References

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, and Jean-François, Rahwan, Iyad Bonnefon. 2018. "The Moral Machine experiment." *Nature* 59-64.

Bauman, W., Christopher, Peter, A. McGraw, and M., Daniel Bertels. 2014. "Revisiting External Validity: Concerns about Trolley Problems and Other Sacrifical Dilemmas in Moral Psychology." *Social and Personality Psychology Compass* 536-554.

Canis, Bill. 2019. *Issues in Autonomous Vehicle Testing and.* CRS Report, Congressional Research Service.

2015. *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey.* A Brief Statistical Summary, Washington, D.C.: National Highway Traffic Safety Administration.

Cunneen, M., M. Mullins, F. Murphy, D. Shannon, I. Furxhi, and C. Ryan. 2020. "Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics." *Cybernetics and Systems* 51 (1): 59-80.

Dubljević, Veljko. 2020. "Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles." *Science and Engineering Ethics* 2461-2472.

Dubljević, Veljko, and Eric Racine. 2014. "The ADC of Moral Judgment: Opening the Black Box of Moral Intuitions With Heuristics About Agents, Deeds, and Consequences." *AJOB Neuroscience* 3-20.

Dubljević, Veljko, George List, Jovan Milojevich, Nirav Ajmeri, William A Bauer, Munindar P Singh, Eleni Bardaka, et al. 2021. "Toward a rational and ethical sociotechnical

system of autonomous vehicles: A novel application of multi-criteria decision analysis." *PLOS ONE.*

Dubljević, Veljko, Sebastian Sattler, and Eric Racine. 2018. "Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment." *PLOS ONE.*

Etienne, H. 2022. "A practical role-based approach for autonomous vehicle moral dilemmas." *Big Data & Society* 1-12.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review* 5-15.

Gamez, Patrick, B., Daniel Shank, Carson Arnold, and Mallory North. 2020. "Artificial virtue: the machine question and perceptions of moral character in artificial moral agents." *AI & SOCIETY* 795-809.

Gopinath, Krishnan, and Gopalakrishnan Narayanamurthy. 2022. "Early bird catches the worm! Meta-analysis of autonomous vehicles adoption – Moderating role of automation level, ownership and culture." *International Journal of Information Management.*

Greene, J. 2013. *Moral Tribes: Emotions, Reason, and The Gap Between Us and Them.* New York: The Penguin Press.

Harris, J. 2020. "The Immoral Machine." *Cambridge Quarterly of Healthcare Ethics* 29: 71–79.

Lin, P. 2016. "Why Ethics Matters for Autonomous Cars." In *Autonomous Driving: Technical, Legal and Social Aspects*, edited by M. Maurer, J.C. Gerdes, B. Lenz and H. Winner, 69-86. Heidelberg: Springer.

Millar, Jason. 2017. "Ethics settings for Autonomous vehicles." In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, by Patrick Lin, Abney Keith and Ryan Jenkins, 20-34. New York: Oxford University Press.

Pflanzer, Michael, Zachary Traylor, B., Joseph Lyons, Veljko Dubljević, and S., Chang Nam. 2022. "Ethics in human–AI teaming: principles and perspectives." *AI and Ethics.*

Sattler, S., V. Dubljevic, and E. Racine. 2022. "Cooperative behavior in the workplace: Empirical evidence from the agent-deed-consequences model of moral judgment." *Frontiers in Psychology* 1-14.

Sharma, Omveer, N. Sahoo, and N. Puhan. 2021. "Recent advances in motion and behavior planning techniques for software architecture of autonomous vehicles: A state-of-the-art survey." *Engineering Applications of Artificial Intelligence.*

Smith, W., Jordan. 2015. "Immersive Virtual Environment Technology to Supplement Environmental Perception, Preference and Behavior Research: A Review with Applications ." *International Journal of Environmental Research and Public Health* 11486-11505.

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

Thomson, Jarvis, Judith. 1985. "The Trolley Problem." *The Yale Law Journal , May, 1985, Vol. 94, No. 6* 1395-1415.

Uhlmann, Luis, Eric, A., David Pizarro, and Daniel Diermeier. 2015. "A Person-Centered Approach to Moral Judgment." *Perspecitves on Psychological Science* 72-81.