Experiments in Moral and Political Philosophy

Edited by
Hugo Viciana, Antonio Gaitán and
Fernando Aguiar

ISBN: 978-1-032-29390-5 (hbk)
ISBN: 978-1-032-29391-2 (pbk)
ISBN: 978-1-003-30142-4 (ebk)

First published 2024

Chapter 15

The Use and Abuse of Moral Preferences in the Ethics of Self-Driving Cars

Norbert Paulo, Leonie Alina Möck, and Lando Kirchmair

The chapter DOI: 10.4324/9781003301424-18

Funded by Digitalization and Technology Research Center of the Bundeswehr

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

15 The Use and Abuse of Moral Preferences in the Ethics of Self-Driving Cars

Norbert Paulo, Leonie Alina Möck, and Lando Kirchmair[*]

Introduction

The technology of self-driving cars is advancing fast. Soon, many vehicles will be on the road that are no longer directed by human beings. This development is not only technologically fascinating but also calls for ethical evaluation and legal regulation. The oft-fatal consequences of car accidents have often been caused by bad reactions of human drivers. These reactions were mostly very fast and affective rather than slow and deliberate. With the technological capabilities of self-driving cars, this is expected to change (e.g., Feldle, 2018, pp. 22–23). For the first time, it will be possible to determine the consequences of unavoidable accidents, because it is possible to program them in advance. The question then is how to program self-driving cars for such accidents.

In this paper, we focus on moral dilemmas, i.e., situations in which, according to all available options, comparable harm occurs, e.g., a group of two or three people is killed because it is not possible to prevent both scenarios (cf. Sinnott-Armstrong 2005), particularly situations in which the car's action "decides" the fate of various humans involved in unavoidable accidents. While Germany took the lead in drafting the first comprehensive law on self-driving cars,[1] the regulation of moral dilemmas still remains ambiguous (for an overview, see Hilgendorf, 2021). The salience of the topic encouraged empirical studies on the moral preferences of individuals about how to handle moral dilemmas caused by self-driving vehicles. The so-called Moral Machine experiment (Awad et al., 2018) stands out due to its large databases of "40 million decisions in ten languages from millions of people in 233 countries and territories." This interrogation of laypeople in order to make life and death decisions, however, stirred a heated debate (see, e.g., Harris, 2020).

In this contribution, we will first give an overview of the design and findings of the Moral Machine experiment (Section "Empirical Research on Moral Preferences in the Ethics of Self-Driving Cars: The Moral Machine Experiment"). On this basis, we will then highlight important methodological (Section "Methodological Criticism: Thought Experiments in Ethics and

DOI: 10.4324/9781003301424-18

Alternative Option") as well as normative (Section "Normative Criticism: Law and Public Morality") criticism. We will show how experimental design choices can (negatively) influence the findings and what is – for normative reasons – important to note when investigating the moral preferences of laypersons empirically. After pointing out the use and abuse of moral preferences in the ethics of self-driving cars, we conclude that, despite potentially misguided methodology and problematic normative issues, empirical moral philosophy has an important role to play when thinking about the regulation of moral dilemmas caused by self-driving cars.

Empirical Research on Moral Preferences in the Ethics of Self-Driving Cars: The Moral Machine Experiment

In the large-scale Moral Machine experiment, Awad et al. gathered evidence about cultures' or societies' moral preferences about accidents involving self-driving cars. Subjects from all around the world were presented with scenarios in which deadly accidents with self-driving cars are unavoidable. The Moral Machine is an online platform, developed at the Massachusetts Institute of Technology, which has been openly available since 2016.[2] Visitors to the platform decide in fictitious dilemma situations about what a self-driving vehicle should do. They must choose between two accident scenarios in which, as a result of their choice, a certain person or group of people (or animals) is killed. For example, imagine this situation: Three pedestrians, two women and a child, are crossing the road. A self-driving car is approaching fast and cannot stop before the pedestrians. If it stays on track, it will kill the child. If it swerves, it will kill the two women. It just cannot save all of them. Should the car stay on track or change course?

The conception of the dilemmas is based on the well-known trolley problem, the original formulation of which goes back to Philippa Foot (1967). It has since been discussed in numerous variations (for an overview, see Bruers & Braeckman, 2014). The Moral Machine experiment uses trolley-like dilemmas in order to derive moral preferences in various societies. Awad et al. (2018, p. 59) assume that the work of ethicists would be "useless" if their recommendations concerning the regulation of decisions about life and death were too different from those of laypersons, which are the eventual subject of the regulation.

In any round of the Moral Machine experiment, participants decide on 13 randomly selected dilemmas. Initially, Awad et al. presented participants with a static 2D rendering of a situation in which there are only two options: either the car kills person/group A or it kills person/group B. Based on certain information about A and B, participants have to decide if the car should kill A or B. From the choices of participants across the 13 randomly composed dilemmas, they distilled preferences into nine categories: (1) how often people are saved compared to pets, (2) what role the "lane keeping" of the vehicle plays in the decision, (3) whether vehicle occupants or pedestrians are saved, (4) what role the law-abidingness of pedestrians plays (saving jaywalkers or the lawful), (5) whether as many lives as possible are saved, as well as preferences related to the categories of (6) age, (7) gender, (8) fitness, and (9) social status.[3]

In each scenario, at least two of the nine factors are included. After having decided whom to kill and whom to spare, participants are asked to provide further information on their choices and to enter some personal information (age, gender, social status, religious background, and political views). The dataset of Awad et al. includes almost 40 million decisions in 10 languages from 233 countries or regions. These decisions come from 2.3 million participants.

The researchers identified nine global trends with regard to the above-mentioned factors in the participants' decisions. For evaluating them, they calculated the average marginal component effect (AMCE) over all included decisions (*N* = 35.2 million), which is the difference between the probability of sparing characters with one attribute (for instance, elderly people) and the probability of sparing the characters with the "opposite" attribute (for instance, children), over the joint distribution of all other attributes. They highlighted three dominant preferences as relevant for a "universal machine ethic" (Awad et al., 2018, p. 63): (1) the tendency to save people before pets (AMCE=0.58, which means that the probability that people are saved is 0.58 higher than the probability that pets are saved), (2) the tendency to save as many lives as possible (0.51), and (3) the tendency to save younger before older lives (0.49).

In addition to these findings, they also observed the preference for sparing the lives of the lawful (0.35) as well as preferences concerning personal characteristics, such as the preference for sparing people with a higher social status (0.35), sparing the lives of fit before unfit people (0.16), and women before men (0.12). They observed (weak) preferences for saving pedestrians over passengers (0.1) and a preference for deciding against vehicle action through swerving (0.06). Finally, Awad et al. pointed out the four most spared characters: baby, little boy, little girl, and pregnant woman.

In addition, the authors investigated individual and cultural variations and correlations in the decision-making patterns using the participants' extended data and geolocation. In the case of individual variations, Awad et al. found no decisive influence of the individual categories on the decision preferences. These variations were deemed negligible for the regulation of self-driving cars.

A different conclusion was drawn with regard to cultural variations. In the course of a cluster analysis of the location and decision data, Awad et al.

compiled three cultural clusters: (1) North America and many European countries (the "Western cluster"), (2) many far Eastern Countries (the "Eastern cluster"), and (3) Latin America and countries with French influence (the "Southern cluster") (2018, p. 61). The authors found partly strong differences in preferences between the clusters in all nine categories. For example, subjects from the Eastern cluster attributed much less weight to sparing younger vs. older people than subjects from the other clusters. They also found correlations between the response behavior and certain cultural and socio-economic aspects such as economic inequality, individualism, rule of law, or gender inequality. For example, the level of social inequality (represented by the Gini coefficient) correlates with the preference to spare people due to their social status. As a conclusion of these findings, the authors argue that policymakers should be, "if not responsive, at least cognizant" of cultural and national differences in moral preferences to ensure acceptance by the respective populations (Awad et al., 2018, p. 61).

Methodological Criticism: Thought Experiments in Ethics and Alternative Options

The Moral Machine experiment has been criticized for methodological as well as for normative reasons (see, e.g., Furey & Hill, 2021; Kochupillai, Lütge, & Poszler, 2020; Nascimento et al., 2019). In this section, we will focus on the methodological criticism of the study before turning to the normative criticism in the next section.

Unfamiliar Thought Experiments

Much of the methodological criticism is actually not particular to the Moral Machine experiment. When the study is criticized for using an unrealistic or simplistic setup modeled after the trolley problem (Dubljević, 2020; Goodall, 2016; Himmelreich, 2018; Nyholm & Smids, 2016; Roff, 2018),[4] this mirrors the general criticism of the use of thought experiments in philosophy. In ethics, thought experiments are often labeled "outlandish," "far-fetched," or "fanciful," and intuitive responses to them are said to be of little or no epistemic value because they are "unfamiliar" (Fried, 2012; Wood, 2011). Designers of thought experiments often stipulate certain unrealistic features to hold in the relevant scenario; for instance, they stipulate that contextual information that in real life would change the moral complexion of tragic choices is irrelevant in the scenario or that the outcomes of all available choices are known with certainty ex-ante. If such unrealistic thought experiments are used to trigger moral intuitions, these seem to be epistemically unreliable because the scenario differs too much from the learning environment that shaped these intuitions. Psychological research with trolley cases in particular seems to lack external validity (Bauman et al., 2014). It is thus argued that ethicists should stick to realistic cases (Wilkes, 1988).

Arguably, the most promising response to such objections against fanciful thought experiments such as the trolley cases draws on their pairwise or sequential use (Wilson, 2016). What is philosophically interesting about thought experiments – especially in ethics[5] – is often not the individual intuitive response to a certain unrealistic scenario, but the contrasting responses to two or more similar yet distinct scenarios. The finding that calls for philosophical attention is that the contrasting responses are stable across large populations. For this kind of convergence of contrasting responses, the unrealistic character of the individual cases is irrelevant, or so it is argued (Greene, 2014; Kamm, 2009; critically Sauer, 2018, Chapter 6; Wood, 2011).

As the Moral Machine experiment used a sequence of unrealistic trolley-like scenarios, the study can be defended with this response, at least in principle. That is, in order for the contrasting-responses defense to work, it might be necessary to make minor changes to the design of the study. Whether or not the contrasting-responses defense (or another defense) ultimately is convincing is a question that concerns almost all uses of thought experiments in ethics. It is not a specific question for the Moral Machine experiment. For the purposes of this paper, this means that, as long as there is no clear answer to the general question regarding the contrasting-responses defense, there is no definitive reason to disregard empirical studies because they use unrealistic scenarios.

Lack of a Third Option

Turning to the Moral Machine experiment, a particularly intriguing methodological critique comes from Yochanan Bigman and Kurt Gray (2020). Their main idea is straightforward: Perhaps some of the results of the Moral Machine experiment are unintended consequences of a flaw in the design of the study. Recall that Awad et al. presented participants with situations with two exhaustive alternatives: either the car kills person/group A, or it kills person/group B. Based on certain information about A and B, the test persons had to decide whom the car should kill.

What that choice design lacks, Bigman and Gray suspected, is a third option, namely one that allows the test persons to treat the potential victims equally. Bigman and Gray tested whether participants' responses differ from those in the Moral Machine experiment when one adds what they call an "equality option": either the car kills person/group A, or it kills person/group B, or it treats the lives of person/group A and B equally.

As Bigman and Gray explain, this equality option can mean that the car simply ignores (or fails to detect) the personal features of A and B.

Bigman and Gray ran three[6] online vignette studies on the ethics of self-driving cars. Of course, they had a much smaller sample size than the Moral Machine experiment. The first study (*N* = 2,352 (the USA and the UK)) tested what participants thought how self-driving vehicles should be programmed. In the first step, the study successfully replicated some of the main findings of the Moral Machine experiment by using a study design similar to the Moral Machine.[7] For example, they find that 96.1% prefer programming that saves children instead of elderly people (only 3.9% prefer to save the elderly instead of the young) and that almost all participants (99.6%) prefer to save more lives rather than fewer.

However, they discover discrepancies with the Moral Machine experiment when changing the study design so that the choice is no longer restricted to two bad options – killing one type of people (A) to save another (B) – but includes a third option ("equality allowed"), namely "having AVs programmed to treat both categories equally (e.g., 'Treat the lives of men and women equally')" (Bigman & Gray, 2020, supplementary material). Many participants seem to prefer this equality option over killing either A or B. For example, Bigman and Gray find that, with that additional option, the preference to save children drops to 22.2% (0.5% prefer to save the elderly), whereas 77.3% prefer the equality option. Similarly, the preference for saving more lives drops to 60.0% (strangely, 0.3% seem to prefer to save fewer people), whereas 39.7% choose the equality option (Bigman and Gray, 2020, see supplementary table 1).

This looks like a quite significant discrepancy with the findings of the Moral Machine experiment. However, Bigman and Gray's first study likely suffers from an unintended influence of the wording of the equality option: it is the only option that does not mention "killing." People might thus (implicitly) assume that choosing the equality option means that no one is killed.

In their second study (*N* = 843 (US)), Bigman and Gray use another formulation of the equality option, which avoids the potential framing problem.[8] In this study, they find that 38.8% would prefer to save the young (over the elderly) and that only (61.1%) would rather treat them equally, which indeed suggests that there was a framing effect in study 1. However, their results in favor of the equality option are nevertheless significant. Also, with the "killing" frame in study 2, 81.6% prefer to save more lives, and the preference for equality goes down to 17.9% (Bigman and Gray, 2020, supplementary table 2).[9]

Summing up, Bigman and Gray find that participants largely prefer the equality option over killing either A or B (with only one exception, law-abidingness). From this, they conclude "that the current [Moral Machine] paradigm is relatively insensitive to preferences for equality" (2020, p. E2).

Response from Awad et al.

In their reply to Bigman and Gray, Awad et al. (2020) point to an unpublished part of the Moral Machine experiment. This part concerns an option in the study. Participants were asked if they want to correct their weighing of the nine preferences resulting from their decisions by moving a slider between the variables (e.g., between human and pet, female and male, etc.). For example, when one's choices in the dilemmatic scenarios (with two options) amount to a preference for killing elderly people rather than children, the slider would be positioned closer to the right end of the scale (children) than to the left end (elderly people). Awad et al. hold that this option to correct one's choices resembles the equality option in Bigman and Gray's second study because test persons were able to correct their choice by moving the slider to the middle position between children and elderly people.

Awad et al. (2020, p. E4) report that, although more than 99% of participants (*N* = 585,531) who were offered the chance to correct their choices did in fact move the slider for at least one dimension, this made no significant difference for some preferences. The strong preferences revealed by the Moral Machine experiment (saving humans, saving more lives, saving younger lives, and saving the law-abiding) seem to remain strong. Participants did not move the slider to the middle position to express their preference for equality between children and elderly people, for example. Awad et al. also find that there are certain factors for which participants did move the slider to the middle position, thereby expressing a preference for equality. However, since these factors concern preferences the Moral Machine experiment identified as weak (for example, the preference for killing men rather than women), they take this as further evidence for their weakness.

Summing up, Awad et al. hold that – with one minor exception[10] – the possibility to express a preference for equality (that was not available in the initial test scenarios with two options) has no significant effect on the results of the Moral Machine experiment.

Possible Understanding of the Equality Option

Focusing on the strong preferences revealed by the Moral Machine experiment (saving humans, saving more lives, saving younger lives, and saving the law-abiding), it is noteworthy that Awad et al. and Bigman and Gray do not find the exact same preferences. Consider the preference for saving more lives. This preference is very strong in the initial two-option scenarios of the Moral Machine experiment. When given the chance to correct this result, many use it to move the slider closer to the middle position, resulting in a somewhat weaker (but still strong) preference for saving more lives. As reported above, Bigman and Gray (2020, supplementary table 2) also find that 81.6% prefer to save more lives, and that (17.9%) opt for equality. But consider the strong preference for saving children over elderly people found in the Moral Machine experiment. Again, when given the chance to correct this result, many use it to move the slider closer to the middle position, resulting in a somewhat weaker (but still strong) preference for saving children. In contrast, Bigman and Gray find that only 38.8% would prefer to save the young. Most test persons (61.1%) would rather treat them equally.

One possible explanation for these discrepancies is that Awad et al.'s results of the slider option might be influenced by a commitment and consistency bias. After all, participants might (implicitly) not want to correct or contradict their previous decisions. This is not so much a criticism of their study because the slider option was not designed as a tool for testing something similar to Bigman and Gray's equality option. As Awad et al.'s reply to Bigman and Gray makes clear, they merely interpreted the data gathered with the slider in terms of an equality option. Nevertheless, the results of Bigman and Gray might indicate that what the Moral Machine experiment investigates really are somewhat forced preferences between killing A and killing B, while many participants might have preferred another alternative option.

Yet, the equality options offered by Awad et al. and Bigman and Gray respectively can be understood in quite different ways. For example, an equality option can be understood in terms of impartiality (on the notion of impartiality, see Jollimore, 2022). This would entail that each person counts the same, i.e., personal characteristics are disregarded. This, together with a general rule to stay on track rather than swerving, seems to be what Bigman and Gray had in mind. However, it is unclear if their participants had the same understanding. Another quite natural understanding of an equality option is the use of random choice. There are yet more possible understandings, of course, but impartiality and random choice seem to be the most obvious ones.

If further studies made transparent to participants that equality means randomization, say, they might well reveal that they in fact do not have a preference for equality thus understood. After all, this option would lead to many situations (roughly 50%) in which more people would be killed although the car could have killed fewer people. Remember that Bigman and Gray's study suggests that almost no one has this preference: only 0.3% report a preference for saving fewer rather than more people (Bigman and Gray, 2020, see supplementary table 2).

It remains open how participants understood Bigman and Gray's equality option and if Awad et al.'s slider really tracked something similar to it. Further research is needed to reveal if there is a preference for a third option and if this is impartiality, randomization, or yet another understanding of equality. Eventually, it is also possible that participants would prefer an option that allows them not to make that kind of decision (and delegate it to experts, say).

Normative Criticism: Law and Public Morality

In this section, we will focus on the normative criticism of the study. As a pronounced critique of the Moral Machine experiment, the moral philosopher John Harris rejects any moral value of the Moral Machine, calling it a "useless" project of huge "naiveté" (2020, p. 73) and defaming Awad et al. as "Moral Machinists" throughout his paper. He considers the main problem of the experiment to be a large-scale trivialization of the relevant issues and accuses Awad et al. (2018) in several aspects for neglecting the required awareness of the relevant moral and legal aspects. Therefore, the experiment would – according to Harris – not contribute to the ethical debate and the foreseeable legal regulation of self-driving cars. While there are many critical perspectives on the Moral Machine experiment, in what follows we will take Harris' critique as a reference point. We do so, because he offers some comprehensible criticism of the Moral Machine experiment, voicing various arguments and concerns in a pointed manner, especially concerning the danger of trivializing legal rights in democratic societies as well as of taking moral matters too lightheartedly. Nevertheless, his full rejection of the study for the below-mentioned reasons is not tenable. We will thus argue why, if carefully designed and executed, experiments like the Moral Machine can indeed inform ethical and legal debates about the regulation of self-driving cars.

The Moral Dimension of the Law

Harris criticizes the Moral Machine scientists for careless handling of the law. He illustrates his criticism by referring to two famous cases of the English legal system and states that moral decisions about the life and death of "innocent road users" could not be debated isolated from the law – "a cognizance of which the Moral Machinists show no evidence whatsoever!" (2020, p. 72). In the first historical landmark case (R v. Dudley and Stephens [1884] 14 QBD 273 DC), two shipwrecked sailors had killed and eaten their cabin boy who had fallen into a coma. Despite acting out of necessity to save their own and a third life, both were charged with murder. Harris blames Awad et al. for taking for granted the permission that a self-driving car could kill innocent people in the case of necessity while "it is not open to the drivers of driverless cars […] automatically to expose other innocent road users to injury or death when the alternative involves any risk to themselves or their machines" (2020, p. 71). With the second example – the case of the separation of the Manchester conjoined twins, killing the weaker one (Re A [2001] 2 WLR 480) – Harris emphasizes the complexity of the legal process which is necessary to deal with life-and-death decisions. "What, absent consent, can be imposed on innocent citizens" is a complicated process in "mature democracies" that requires a broad discourse and inclusion of all relevant aspects. Therefore, to him, "to settle in advance the legal and ethical ramifications of any deaths resulting from the programming of the vehicles […] is naïveté of heroic proportions!" (2020, p. 73). In our view, Harris' first major point, that moral preferences by laypersons must not violate the law, is problematic for several reasons.

First of all, the law is not static but open to changes orchestrated by a democratic law-making process. This process might well be informed by the preferences of the public; the democratic law-making process does not require specifically moral preferences. In principle, any preference supported by a majority might become law, if they are not unconstitutional. Most importantly, this means that, in constitutional democracies, new laws must not violate constitutionally guaranteed fundamental rights. This, however, is only the legal perspective.

A second point worth noting is that moral preferences must not necessarily match the law. While it might be moral to comply with the law, it is not automatically amoral or immoral to have preferences that violate the law. Similarly, asking for preferences is not amoral or immoral merely because the preferences one finds stand in conflict with the law. Just as there are laws that are amoral (e.g., traffic laws stipulating on which side of the road cars must drive), there are certainly also immoral laws, i.e., laws that are immoral but still count as law (think of some contemporary migration laws in contrast to particularly severe "Nazi laws" that arguably do not count as law). Thus, we need reasons as to why some preferences are immoral. Simply stipulating that they would be immoral because of a conflict with the law is not convincing. Without further reasons, it is furthermore circular to say that some preferences are immoral because they might violate the law. This is so because one could also say that laws in conflict with moral preferences should be changed.

Third, there are no good reasons for categorically stating that collecting preferences from laypersons for the sake of (legal) regulation would be meaningless. It is questionable whether this would be different if such preferences were not general preferences, but moral ones (and thus might be particularly sensitive in terms of the process of collecting preferences). In other words, where is the difference between finding the majority has a particular preference in contrast to the majority having a *moral* preference such as in cases of moral dilemmas for saving the young instead of the elderly?

Fourth, all the well-known arguments in favor of the conceptual separation of law and morality that basically aim at shielding the (process of) law (-making) from external influences (see, e.g., the classical essay by Hart, (1958); for an overview, Ratnapala, 2017, Chapter 8) can also be turned upside down in order to shield morality from the law. In this vein, any conceptual understanding as to what morality is should not be influenced by anything only the law might prescribe.

The Legal Notion of Necessity

The specific critique that even in the "shipwrecked sailors" case it was illegal to kill the innocent cabin boy in order to save the lives of three other shipwrecked sailors is also problematic. This is similarly true for the argument that in the "Manchester twins" case it was legal to separate the twins and thereby kill the weaker (in terms of survival chances after separation) twin in order to save the stronger twin only after a complicated and lengthy legal process of a mature democracy adhering to the rule of law. Relying on (English) legal terms, Harris states that there is no "defense of necessity to charges of murder" (2020, p. 72).

The example of the English law as a general argument against the necessity case is problematic. First, Harris' general assessment is not entirely correct anymore, because the "Manchester twins" case somewhat departed from the long-standing precedent that there is no necessity defense against murder charges (just see Santoni de Sio, 2017, p. 416). Second, a slightly different image of the circumstances and potential arguments arises when we turn toward the German criminal law doctrine. Section 32 of the German Criminal Code (self-defense), which allows the killing of a person in order to save oneself, and Section 33 (excessive self-defense), allowing to kill a person in order to save someone else, both require an attack as self-defense according to Section 32 para 2 ("any defensive action which is necessary to avert a present unlawful attack on oneself or another"). Therefore, these sections will be rarely relevant in moral dilemmas with self-driving cars. There are, however, also Section 34 (necessity as justification) and Section 35 (necessity as defense). In emergency situations (in case two lives are in danger but none of the endangered persons engages in an illegal attack endangering the life of the other person), the sacrifice of human life is not justifiable. However, an act resulting in saving person A and killing person B might be – according to a progressive view – justifiable out of necessity or it might be excusable (cf. Hilgendorf, 2018, pp. 60–70). Consider the "Carneades" case, much like the case of the shipwrecked sailors, which is the classic example in German criminal law doctrine. This scenario goes back to the time of Carneades of Cyrene (2nd century BC): two shipwrecked sailors see a plank that can only carry one of them. One of them reaches the plank first; when the other sailor also reaches the plank, he is about to drown. He pushes the first sailor off the plank, thus causing him to drown. Is the surviving sailor to blame or did he merely act in self-defense? Following the German doctrine, the action of the shipwrecked sailors killing the cabin boy in order to save themselves has to be excusable because if their behavior would be forbidden the law would command the sailors to sacrifice themselves. As this is too much to ask from individuals, in Germany such a behavior is, in legal terms, considered "excusable" – i.e., it remains illegal but goes unpunished (for a general overview, see Dreier, 2007).

The Difference Between Regulating Moral Emergencies and Moral Dilemmas

Here is another problem with Harris' argument. Both cases referred to by Harris were such that individuals acted in situations of immediate emergency calling for rapid decisions. The regulation of self-driving cars is relevantly different. It does not concern an individual decision, but the general regulation of situations that very likely will occur once the technology of self-driving cars is sufficiently advanced.

It is thus important to distinguish the regulation of moral emergencies from moral dilemmas in traffic accidents involving self-driving cars. The technological innovation of self-driving cars and their computational capacity confront us with a new situation. Computational capacity outstrips human capacities in reacting during accidents, which offers new potential for regulation. So far, this potential was nonexistent due to the cognitive limitations of humans. In traffic accidents, humans mostly react fast and affectively instead of thinking through all possible options. While we might still also say that moral dilemmas of self-driving cars should be solved by given legal provisions, we might also acknowledge that this genuinely new situation asks for new regulation. Consider the example of sacrificing person/group B in order to save person/group A. According to a conservative opinion in German criminal law, such a sacrifice would only be excusable if the person whose life is at stake takes the decision to kill another person to save her life. For if we do not change the current legal situation (at least according to the prevailing conservative opinion in Germany, for instance), we must realize that we decide for letting A *and* B die (instead of killing one person/group and saving another person/group) for the sake of the law remaining "blameless." This seems like a high price to pay because car traffic – in contrast to shipwrecked sailors – is a daily business.

Moreover, there is the duty of states according to the European Court of Human Rights to protect persons from (avoidable) dangers. With the introduction of the technology of self-driving cars, the number of traffic accidents is arguably reducible to a significant extent (Eisenberger, 2017,

p. 102). This logic might be extended to the remaining accidents that are not avoidable. For these moral dilemmas, the obligation of states to protect lives might lead to the obligation to choose either to save A or to kill B instead of letting both, A *and* B, die (cf. Kirchmair forthcoming).

The Concept of Morality

The second major claim Harris makes in his paper is that the Moral Machine experiment lacks any sound concept of morality. The problem, Harris notes, is less the attempt of the "Moral Machinists" to take public morality as the foundation for legal and ethical frameworks, but mainly the reductionist image they draw of public morality, ignoring its embeddedness within a historical and cultural process. Public morality, Harris says, "has evolved over a lengthy period, often painfully; informed by history, art, literature, culture, personal experience, and much more" (2020, p. 76). In contrast, Awad et al. simply consider the participants' preferences on the Moral Machine platform as expressions of public morality, or so he claims:

> Majorities are not necessarily right; neither science nor ethics is produced by casting votes for particular 'answers'; happy though such a possibility might seem to some! The Moral Machinists are proposing the moral equivalent of deciding whether the world is flat by finding out what people would prefer the answer to be.
>
> (Harris, 2020, p. 74)

Following Harris, Awad et al. aim to replace a reflected discourse on complex and controversial ethical problems with simple yes-or-no questions to uninformed individuals. For Harris, this methodology is highly unreliable. "'Public morality', as they crudely and mistakenly understand it, requires only ill informed, unconsidered preferences, given instantly and thoughtlessly, as if playing a computer game!" (2020, p. 76). Therefore, he considers the experiment "amoral and indeed immoral" (2020, p. 78) by promoting a simplified and wrong impression of public morality that consequently eliminates the distinction between moral judgments and personal prejudices. As Bonnefon explains, the idea behind the Moral Machine experiment was not to decide what ought to be done, all things considered, by polling moral preferences. Referring to the German Ethics Commission, he writes: "there is nothing wrong with trusting well-informed specialists, but it is unfortunate that citizens were not given an opportunity to voice their preferences, especially when the specialists disagreed" (Bonnefon, 2021, p. 71).[11]

It is of course a grand question to ask what (public) morality is. The critique, however, that morality cannot be found with an experiment like the

Moral Machine experiment hinges as much upon answering this question as the experiment itself. Is morality only to be found in the "ivory tower" of ethical theory building or is it (also) connected to what (the majority of) laypersons consider(s) to be the right thing to do? If the latter has to play a role, the study design in order to find morally relevant preferences becomes crucial. While it is correct to ask for high methodological standards when investigating preferences, in particular moral preferences, it is too easy to wipe off any experiment aiming to shed light on public morality.

Similarly, the role of the majority in public morality is a complex issue. It seems that when avoiding cases in which the majority is directed against a specifiable minority (e.g., if a population with 70% elderly persons is asked whether either the elderly or the young should be saved and the vote goes in favor of the elderly), the judgment of the majority of a population as well has a decisive role to play in finding public morality, for if this would not be the case we could speak of the tyranny of the (moral) few. However, if the majority prefers sacrificing some minority for dubious reasons (let self-driving cars kill a specific minority group first), we need to be able to dismiss their preferences somehow.

Moreover, it does not suffice to stipulate the irrelevance of "what the people think" for what is morally right. Some political philosophers take this to be of high importance for questions of justice. David Miller, for instance, holds that any theory of justice is "to be tested, in part, by its correspondence with evidence concerning everyday beliefs about justice" (1999, p. 51). Also, neglecting this relevance might force one to endorse a version of metaethical universalism, i.e., the idea that there is a single true morality that applies to all individuals and groups. As Thomas Pölzler has recently argued, metaethical universalism is way more controversial than is commonly thought, and metaethical relativism is perhaps the more convincing position. In contrast to universalism, relativism holds that "the truth or falsity of moral judgments depends on the beliefs, traditions, practices, sentiments etc. of individuals or groups" (Pölzler, 2021, p. 834). Such a position would arguably be easier to combine with a certain normative relevance of public preferences. This seems also true for yet another position in the literature, according to which the actual moral preferences of the people are so diverse – and likely to remain so – that we should aim for a political solution rather than a moral one (Brändle & Schmidt, 2021; Himmelreich, 2018, pp. 675–676; Rodríguez-Alcázar, Bermejo-Luque, & Molina-Pérez, 2020).

Finally, saying that majorities don't necessarily make morally right judgments, a point that Harris underlines with the example of the former popularity of a "certain 'Bohemian corporal[,s]'" opinion, doesn't show that they conversely most likely arrive at wrong moral judgments. Certainly, majorities are not a guarantee for morality but so aren't judgments of individuals, even those of moral philosophers. Harris seems to endorse a concept of morality that requires a person to possess a certain level of expertise or at least being informed in order to make moral judgments. He criticizes the Moral Machine experiment for considering "ill informed, unconsidered preferences" and throughout his paper avoids speaking of *moral* intuitions or *moral* preferences in the context of the Moral Machine experiment. Yet, recent empirical research on the moral intuitions of professional moral philosophers has challenged the idea that there is such a thing as moral expertise in the sense that the moral intuitions of moral philosophers are more reliable than those of laypersons (for an overview, see Horvath & Koch, 2020; Paulo, 2020, pp. 345–356). As Eric Schwitzgebel and Fiery Cushman put it in their well-known study of philosophical expertise,

> "if there is a level of philosophical expertise that reduces the influence of factors such as order and frame upon one's moral judgments, we have yet to find empirical evidence of it."
>
> (2015, 136)

Conclusion and Outlook

In this contribution, we have argued for the potential of empirical moral philosophy in the context of the regulation of self-driving cars. For this, we focused on the use and abuse of capturing the moral preferences of the public and including these into the regulatory process.

The Moral Machine experiment, probably the most prominent empirical study in this field due to its large dataset, was used as an example of a study that collected evidence on public moral preferences for the sake of programming self-driving vehicles. After a comprehensive presentation of the study's aim and methodology, criticism of the study has been discussed and partly refuted.

On the one hand, critical points concerning the methodology (or the "right" way of investigating the public's moral preferences) were discussed. We found that, while Awad et al.'s approach of using thought experiments for their experiment can be defended, there is some legitimate criticism that has to be taken seriously. As Bigman and Gray have shown with the introduction of a third option – the equality option – in their own studies, the Moral Machine experiment might suffer from the danger of distorting the results through the experimental setup, namely by forcing participants to decide between killing group A and group B. This limited choice in the Moral Machine experiment design might hamper capturing the real preferences of people in dilemma situations. Although rejected by Awad et al. by pointing to their own version of an equality option, this bug will have to be considered in new experiments if the point is to actually discover the moral preferences of individuals (instead of forcing them to take political decisions of how to regulate dilemma situations). Also, it turned out that conceptual problems concerning the scope of interpreting the preference for equality in both studies point toward the necessity of further research on this option. The notion of equality simply is too broad in order to serve as an unambiguous third alternative without blurring the actual moral preferences of test persons.

On the other hand, we have shown that, whereas methodological criticisms concerning the Moral Machine experiment raise important questions, arguments for a total rejection of this (or a similar) study for normative reasons don't hold. As a response to Harris' *Immoral Machine*, we have argued that neither from a legal nor from an ethical perspective, there is a point of concern that convincingly undermines the raison d'être of the Moral Machine experiment. Still, it has indeed to be stressed that investigating public morality is a sensitive goal that requires critical evaluation of the methodological approaches throughout the process. Sensible ways of accounting for both public morality and traditional moral theory have been suggested (see, e.g., Savulescu, Gyngell, & Kahane, 2020; Paulo, 2023).

As a result, it can be stated that the findings of the Moral Machine experiment are an impressive collection of data that has indeed contributed to the ethical and legal debate of how to regulate moral dilemmas caused by self-driving cars. Future empirical research in the field can continue this course. While the methodological limits of the Moral Machine experiment have to be acknowledged, it is nevertheless important to consider public moral preferences in the ethics of self-driving cars.

Finally, it is important to consider the context when discussing the ethics of self-driving cars. While from today's perspective, self-driving cars promise to save lives (as currently most of the lethal car accidents are caused by human errors), this strong gain in safety might decrease in the wake of time. Once self-driving cars will be "the new normal," and human drivers the exception, the assessment standard will change. In the future, accidents due to human errors might be forgotten and the remaining causes of accidents will be the dominant issue. Pointedly expressed, what could be an acceptable and reserved programming for self-driving cars in 2023 might not be acceptable anymore in 2033. With this shift, moral preferences on how to regulate dilemma situations might also change. The likely changing ratio from a majority of human-driven cars in 2023 and a likely majority of self-driving cars in the mid-term future might thus also require different rules. While a mostly human-driven car environment might be best guided by currently dominant rules, an almost only self-driven car environment might work with a (almost) new set of rules, morally and legally. In other words, the ethics of self-driving cars are context dependent and cannot be considered isolated. Approaches of empirical moral philosophy working on these moral questions should accompany this process by addressing pressing questions at hand. This means that empirical work on moral preferences should focus on the pressing challenges that are likely about to change considering the current state of affairs.

Acknowledgments

This research is part of the project EMERGENCY-VRD and funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr which we gratefully acknowledge. dtec.bw is funded by the European Union – NextGenerationEU. Thanks also to Yochanan Bigman, Thomas Pölzler, Sebastian Krempelmeier, and an anonymous reviewer who provided comments on earlier versions of this chapter.

Notes

[*] All authors of this chapter contributed equally

[1] The provision on collision avoidance systems (and dilemma problems) in § 1e para. 2 No. 2 of the German "Gesetz zum autonomen Fahren" of July 2021 (Federal Law Gazette I 2021, p. 3089, 3108), reads as follows: Motor vehicles with an autonomous driving function must have an accident avoidance system that

> "(a) is designed to avoid and reduce harm,
> (b) in the event of unavoidable alternative harm to different legal interests, takes into account the importance of the legal interests, with the protection of human life having the highest priority; and
> (c) in the case of unavoidable alternative harm to human life, does not provide for further weighting on the basis of personal characteristics." (Translation: Lando Kirchmair)

[2] For more information on the setup and working of the Moral Machine, see Bonnefon (2021).

[3] In addition, further categories were integrated that, according to Awad et al., cannot be assigned to one of the nine factors (e.g., pregnant women, doctors or criminals).

[4] Others have defended the use of trolley cases in the debate concerning accidents of autonomous vehicles, see, e.g., Keeling (2020) and Wolkenstein (2018). For a detailed discussion of the actual philosophical trolley problem in the ethics of self-driving cars, see Kamm (2020).

[5] On the types and functions of thought experiments in ethics, see Pölzler and Paulo (2021).

[6] The third of these is not directly relevant for the purposes of this paper, which is why we only mention studies 1 and 2.

[7] They call this the "forced inequality" condition, which is described as follows: "participants were given two options to choose from: killing one type of people to save another (e.g., "kill men and save women") or vice versa (e.g., "kill women and save men)" (Bigman & Gray, 2020, supplementary material).

[8] As Bigman and Gray explain, "The procedure was identical to the 'Equality Allowed' condition from Study 1 with one change: The 'Treat Equally' option was framed 'To decide who to kill and who to save without considering

whether it is XXX or YYY'. For example: 'To decide who to kill and who to save without considering whether it is a man or a woman'" (2020, supplementary material).

[9] In Bigman and Gray's supplementary table 2, the numbers concerning fewer vs. many are inadvertently reversed, as Yochanan Bigman has confirmed to us in personal communication.

[10] This exception concerns the social status of the potential victim. Whereas the responses to the dilemmatic scenarios revealed a preference for saving people with high social status, test persons largely corrected this result using the slider (Awad et al., 2020, p. E4).

[11] He also posits that the Moral Machine experiment included controversial categories such as social status and bodily fitness in order to make clear that "people shouldn't blindly follow our results when programming self-driving cars" (Bonnefon, 2021, p. 47).

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2020). Reply to: Life and death decisions of autonomous vehicles. *Nature*, *579*(7797), E3–E5. https://doi.org/10.1038/s41586-020-1988-3

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, *579*(7797), E1–E2. https://doi.org/10.1038/s41586-020-1987-4

Bonnefon, J.-F. (2021). *The Car That Knew Too Much: Can a Machine Be Moral?* Cambridge, MA: MIT Press.

Brändle, C., & Schmidt, M. W. (2021). Autonomous driving and public reason: A Rawlsian approach. *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00468-1

Bruers, S., & Braeckman, J. (2014). A review and systematization of the trolley problem. *Philosophia*, *42*(2), 251–269. https://doi.org/10.1007/s11406-013-9507-5

Dreier, H. (2007). Grenzen des Tötungsverbotes – Teil 1. *JuristenZeitung*, *62*(6), 261–270. https://doi.org/10.1628/002268807780282749

Dubljević, V. (2020). Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics*, *26*(5), 2461–2472. https://doi.org/10.1007/s11948-020-00242-0

Eisenberger, I. (2017). Das Trolley-Problem im Spannungsfeld autonomer Fahrzeuge: Lösungsstrategien grundrechtlich betrachtet. In I. Eisenberger, K. Lachmayer, & G. Eisenberger (Eds.), *Autonomes Fahren und Recht* (pp. 91–107). Wien: MANZ Verlag.

Feldle, J. (2018). *Notstandsalgorithmen: Dilemmata im automatisierten Straßenverkehr*. Baden-Baden: Nomos.

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

Fried, B. H. (2012). What does matter? The case for killing the trolley problem (Or Letting It Die). *The Philosophical Quarterly*, *62*(248), 505–529. https://doi.org/10.1111/j.1467-9213.2012.00061.x

Furey, H., & Hill, S. (2021). MIT's moral machine project is a psychological roadblock to self-driving cars. *AI and Ethics*, *1*(2), 151–155. https://doi.org/10.1007/s43681-020-00018-z

Goodall, N. J. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, *30*(8), 810–821. https://doi.org/10.1080/08839514.2016.1229922

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) Science matters for ethics. *Ethics*, *124*(4), 695–726.

Harris, J. (2020). The Immoral Machine. *Cambridge Quarterly of Healthcare Ethics*, *29*(1), 71–79. https://doi.org/10.1017/S096318011900080X

Hart, H. L. A. (1958). Positivism and the separation of law and morals. *Harvard Law Review*, *71*(4), 593–629. https://doi.org/10.2307/1338225

Hilgendorf, E. (2018). The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems. In E. Hilgendorf & J. Feldle (Eds.), *Digitization and the Law* (pp. 57–90). Baden-Baden: Nomos.

Hilgendorf, E. (2021). Straßenverkehrsrecht der Zukunft. *JuristenZeitung, 76*(9), 444–454. https://doi.org/10.1628/jz-2021-0145

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, *21*(3), 669–684. https://doi.org/10.1007/s10677-018-9896-4

Horvath, J., & Koch, S. (2020). Experimental philosophy and the method of cases. *Philosophy Compass*, e12716. https://doi.org/10.1111/phc3.12716

Jollimore, T. (2022). Impartiality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2022/entries/impartiality/

Kamm, F. M. (2009). Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs*, *37*(4), 330–345.

Kamm, F. M. (2020). The use and abuse of the trolley problem: Self-Driving cars, medical treatments, and the distribution of harm. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 79–108). New York: Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0003

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293–307. https://doi.org/10.1007/s11948-019-00096-1

Kochupillai, M., Lütge, C., & Poszler, F. (2020). Programming away human rights and responsibilities? "The Moral Machine Experiment" and the need for a more "Humane" AV Future. *NanoEthics, 14*(3), 285–299. https://doi.org/10.1007/s11569-020-00374-4

Kirchmair, L. (forthcoming). How to Regulate Moral Dilemmas Involving Self-Driving Cars: The German Act on Autonomous Driving 2021, the Trolley Problem, and the Search for a Role Model. *German Law Journal*.

Miller, D. (1999). *Principles of Social Justice*. Cambridge, MA: Harvard University Press.

Nascimento, A., Vismari, L., Queiroz, A. C., Cugnasca, P., Camargo, J., & Almeida, J. (2019). The moral machine: Is it moral? In A. Romanovsky, E. Troubitsyna, I.

Gashi, E. Schoitsch, & F. Bitsch (Eds.), *Computer Safety, Reliability, and Security* (pp. 405–410). Cham, Switzerland: Springer.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

Paulo, N. (2020). Romantisierte Intuitionen? Die Kritik der experimentellen Philosophie am Überlegungsgleichgewicht. In N. Paulo & J. C. Bublitz (Eds.), *Empirische Ethik: Grundlagentexte aus Psychologie und Philosophie* (pp. 323–357). Berlin: Suhrkamp.

Paulo, N. (2023). The Trolley Problem in the Ethics of Autonomous Vehicles. The Philosophical Quarterly. https://doi.org/10.1093/pq/pqad051

Pölzler, T. (2021). The relativistic car: Applying Metaethics to the debate about self-driving vehicles. *Ethical Theory and Moral Practice*. https://doi.org/10.1007/s10677-021-10190-8

Pölzler, T., & Paulo, N. (2021). Thought experiments and experimental ethics. *Inquiry*. https://doi.org/10.1080/0020174X.2021.1916218

Ratnapala, S. (2017). *Jurisprudence* (3rd ed.). Cambridge, United Kingdom: Cambridge University Press.

Rodríguez-Alcázar, J., Bermejo-Luque, L., & Molina-Pérez, A. (2020). Do automated vehicles face moral dilemmas? A plea for a political approach. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00432-5

Roff, H. M. (2018). *The Folly of Trolleys: Ethical Challenges and Autonomous Vehicles* [Brookings Institute]. Retrieved from https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/

Santoni de Sio, F. (2017). Killing by Autonomous Vehicles and the Legal Doctrine of Necessity. *Ethical Theory and Moral Practice*, *20*(2), 411–429. https://doi.org/10.1007/s10677-017-9780-7

Sauer, H. (2018). *Debunking Arguments in Ethics*. New York: Cambridge University Press.

Savulescu, J., Gyngell, C., & Kahane, G. (2020). Collective Reflective Equilibrium in Practice (CREP) and controversial novel technologies. *Bioethics*. https://doi.org/10.1111/bioe.12869

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141*, 127–137.

Sinnott-Armstrong, W. (2005). Moral Dilemmas. *Encyclopedia of Philosophy*. https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/moral-dilemmas.

Wilkes, K. V. (1988). *Real People: Personal Identity without Thought Experiments*. Oxford, New York: Oxford University Press.

Wilson, J. (2016). Internal and external validity in thought experiments. *Proceedings of the Aristotelian Society*, *116*(2), 127–152.

Wolkenstein, A. (2018). What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology, 20*(3), 163–173. https://doi.org/10.1007/s10676-018-9456-6

Wood, A. (2011). Humanity as end in itself. In S. Scheffler (Ed.), *On What Matters* (Vol. 2, pp. 58–82). Oxford; New York: Oxford University Press.